THE UNIVERSITY OF ALBERTA

AN ALGORITHM FOR FINDING NATURAL CLUSTERS

by

J. Alan George

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

SEPTEMBER, 1966

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled AN ALGORITHM FOR FINDING NATURAL CLUSTERS submitted by J. Alan George in partial fulfilment of the requirements for the degree of Master of Science.

ABSTRACT

The objectives and applications of numerical classification are reviewed. The relation between similarity and distance in the classification space is clarified. Some published methods for finding clusters are described and evaluated.

A new clustering algorithm is described. The similarities of the population are first ranked in order of decreasing size. Clusters are begun using the available pair of points whose similarity is highest on the list. Possible additions are sequentially considered by searching down the list of similarities until a similarity between a point outside the cluster and one within the cluster is encountered. Admission or rejection is based on average and single linkage criteria. Rejection of a point causes termination of additions to a cluster. The procedure is repeated to show the clusters at a number of different resolutions.

ACKNOWLEDGEMENTS

My sincere appreciation to Professor J.W. Carmichael and to Professor R.S. Julius for their invaluable assistance and guidance in the preparation of this thesis.

Also, my thanks to Dr. D.B. Scott, Head of the Department of Computing Science, University of Alberta, for providing the facilities necessary to carry out this research and to the National Research Council of Canada for providing financial assistance.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Classification

It is difficult, if not impossible, to discuss the objects of any study, or to examine the relationships between these objects unless the objects are labelled or marked in some way and are given some physical or conceptual arrangement or ordering. The term 'classification' indicates such an arrangement or refers to the act of creating such an arrangement.

The subject which deals with the naming and arranging of things is called taxonomy. A 'taxonomic' classification is one in which the labeling system and the classification are not independent. That is, the labeling of the classified objects depends upon the classification rather than merely being an arbitrary assignment. Sokal and Sneath (1963) coined the convenient descriptive phrase 'operational taxonomic unit' (OTU) to denote the objects of a taxonomic study. We will use the term 'OTU' as a general term to denote one of a set of elements to be classified, whether or not the classification being carried out is 'taxonomic' in the above sense.

Many definitions of classification have been proposed. Webster's Dictionary defines classification as "a systematic arrangement in groups or categories according to some established criteria", or "the act or process of classifying". We will use a more general definition of classification as "any physical or conceptual arrangement of a set of OTU's". Hence, by this unrestrictive definition, an arbitrary listing of the labels of the OTU's under consideration would be a classification, although one of little utility.

1.2 Purposes of Classification

The nature and/or success of a classification procedure depends to a large degree upon its purpose. However, if that purpose is too restrictive, the resulting classification will be a special-purpose one which has little information or utility for any other purpose. Such a classification, created on the basis of a single or very few attributes, is often described as 'arbitrary'. For example, we can divide flowers into groups or subsets on the basis of their color, but such a classification tells us nothing about any other attributes of flowers. The classification has practically no predictive value for other properties. A general purpose classification is one intended to be of use for the widest variety of

endeavors, and our aim is to produce this type of classification. From the above discussion, it appears necessary that such a classification should be based on as large a number of attributes as possible. Indeed, any attribute in which a scientist might conceivably be interested ought to be included. Such a classification has been referred to as 'natural' by numerical taxonomists.

A preliminary to making any classification is the practical problem which Fairthorne (1961) refers to (in the library classification context) as "marking and parking". That is, giving names to the OTU's or labeling them in some way and finding some physical arrangement so that they may be located and retrieved for examination. It is perhaps questionable whether, in most cases, this process is scientific; classification labels, like any others, are not models imaging the things they stand for but are only symbols which lack, by themselves, any characteristics of the objects they denote.[1] This process of "marking and parking" is, however, a mandatory first step in a taxonomic study.

---

[1] For example, Fairthorne (1961), points out that it is our interpretation of the symbol (!) which makes it appear surprised.

When the practical problem outlined above has been overcome, we are free to proceed with the main purpose of classification. That is, to indicate and summarize some relationships existing between the OTU's involved in the study. When one realizes that there are 1770 different possible pairs of 60 OTU's and that we may be interested in many relationships between each pair, it becomes obvious how little one can gain by inspection of a listing or matrix of the relations. Clearly, for studies involving even a moderate number of OTU's and/or relations, some reduction of the data is necessary before even the gross features of the data become noticeable. That is, we need to simplify the information so that it is conceptually manageable. Partitioning or 'clustering' the OTU's into a number of groups whose characteristics within the groups are relatively constant is one way of achieving some economy of memory. If there is high constancy and mutual correlation of characters (i.e. if the classification space is 'unevenly filled'), such a grouping will have a high predictive value. Another device commonly used in conjunction with the above is that of a nested hierarchy. This is the practice of combining a number of groups into fewer, larger ones of higher rank. Care must be exercised, however, when utilizing this concept since useful hierarchies

can only result from certain types of distributions. Although hierarchical classifications may be devised for uniform or random distributions, the classification is not of much use. This is a weakness of classification procedures which yield a dendogram regardless of its suitability for the relationships. In order to form hierarchies, one needs a 'clumped' distribution as in Figure 1 below, where the points represent the OTU's and the distances between the points represent the relationships in which we are interested. The dotted lines indicate the hierarchical levels:

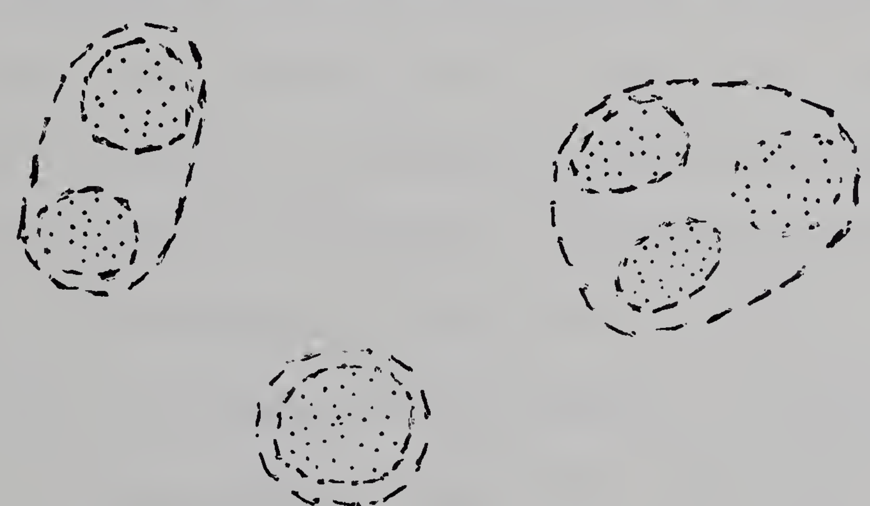


Figure 1a

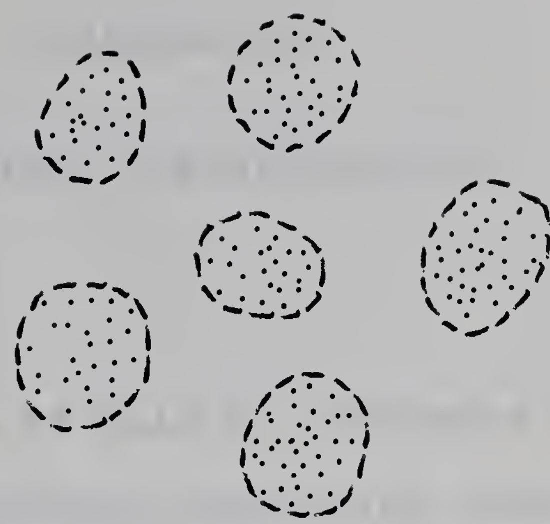


Figure 1b

For a satisfactory hierarchical classification, the OTU's and taxa must be clustered at every level. The distribution in Figure 1b can be grouped at only one level.

[illegible]

[illegible]

CHAPTER II

NUMERICAL CLASSIFICATION

2.1 Definition

In this thesis, we will be concerned only with the problems of classification, neglecting those of nomenclature. With the advent of large-scale, high-speed digital computers, interest in quantitative or numerical taxonomy has increased. It is hoped that quantitative methods can extend the scope of the intuitive interpretations of classical taxonomists and clarify the basis on which intuitive classifications are made.

By numerical classification we mean the quantitative evaluation and expression of relationships between OTU's. According to Sokal and Sneath (1963), the prime aims of numerical classification are repeatability and objectivity. Using numerical methods, different scientists should obtain the same classification for a set of OTU's, given the same data and the objective of general utility. This would do much to eliminate the individual and poorly defined approach which has characterized biological classification up to now. That is, it would make classification more objective by defining criteria and

a standard procedure for grouping the OTU's. By using many characters or attributes, and a standard methodology, we should be able to produce a classification which would be stable to the extent that it would be changed only by new discoveries rather than individual predjudices.

2.2 Relational Coefficients

The first step in the compression of data which most numerical procedures have followed to date is the determination of some single value for each pair of OTU's which reflects their overall relationship, whether it be proximity, similarity or whatever. This value may be based on the ordinary correlation coefficient, so called similarity coefficients, and so on. We will refer to all of them as similarity coefficients or just 'similarities', and a 'high similarity' will imply a high degree of 'sameness' between the OTU's. These coefficients are used as the input data for the various further steps in the classification.

Three general types of similarity measures have been suggested in the literature. These measures can be classified as coefficients of (1) association, (2) correlation, and (3) distance.

Coefficients of association are generally restricted to characters subdivided into only two states (commonly called features). These coefficients vary considerably as

to their formulation but are all based on a 2×2 table arrangement of the data for each pair of OTU's. The four cells of the table have in them the following values: (a) the number of features possessed by the first OTU, but not the second, (b) the number of features possessed by the second, but not by the first, (c) the number of features possessed by both of the OTU's, and (d) the number of features possessed by neither of the OTU's. A typical example of one of the many coefficients of association is the following:

$$S_{ij} = n_{ij}/\ n$$

where S_{ij} : the similarity between OTU i and OTU j.

n_{ij} : number of features OTU i and OTU j have in common.

n : total number of features being considered.

The second type of relational coefficient or similarity coefficient which has been employed is the ordinary Pearson correlation coefficient of the product moment type, permitting characters to be divided into more than two states. It has the following form:

$$r_{jk} = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{(\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{n}(X_{ik} - \bar{X}_k)^2)^{\frac{1}{2}}}$$

where X_{ij} = character state value of character i of OTU j.

$\bar{X}_j$ = mean of character state values of OTU j.

The third type of relational coefficient is based on a geometric model. The value of the coefficient of similarity between a pair of OTU's with n attributes or characters is a function of their distance apart in n-space whose coordinate axes are the characters. This type of measure seems appealing since people tend to think of 'similar things' as geometrically close; that is, similarity and distance are considered to be complements. Various distance measures of similarity have been proposed (Sokal and Sneath 1963), the most useful one being the ordinary unweighted, normalized (in some way) Euclidean distance:

$$D_{ij} = \left[\sum_{k=1}^{n}(X_{ik} - X_{jk})^2\right]^{\frac{1}{2}}/\max D$$

where D_{ij} = the normalized Euclidean distance between OTU's i and j.

The term in the denominator (the normalization factor) deserves some consideration. If we use as the normalization factor the maximum distance that <u>exists</u> between any pair of points (OTU's), our distances will range from some small non-negative number to unity, and the most distant pair of points (taking $D_{ij} = 1 - S_{ij}$ or $S_{ij} = 1 - D_{ij}$) will have a similarity of zero. An alternative is to use the maximum possible distance between the points or OTU's as the normalization factor. This would guarantee that a similarity of zero really did imply that the points were as far apart as they could be with respect to every attribute (coordinate). They would then be hyper-diagonally opposite in the attribute space. This would, in effect, define the classification space as a hyper-rectangular solid rather than a hyper-sphere. The latter alternative seems preferable because it is the usual kind of space we use for graphs and models.

2.3 <u>The General Problem</u>

Using the last alternative (above), we will calculate the distances (1-similarities) between the OTU's as if the attribute values were coordinates determining points in a space with dimensionality determined by the number of attributes. The proportion of the attribute space is determined by the ranges of the attribute values. When

given numerous OUT's there are too many relations (possible pairs of OTU's) to be conceptually manageable. When there are more than three attributes, we can not visualize the configuration of points as they are arranged in the attribute space. That is, we no longer have a physical analogue of the attribute space. The problem, then, when one or both of the above situations exist, is to find a parsimonious descriptor of the relations between the OTU's which loses as little information as possible about the relations, but allows visualization. It is felt that some type of distance model best demonstrates these relations since the concept of similarity or dissimilarity for most people seems to have some inherent spatial qualities (Carmichael et al. 1965). Using such a model, i.e. points in Euclidean space with $D_{ij} = 1 - S_{ij}$ (or some variation $D_{ij} = f(S_{ij})$), poses no problem as long as few OTU's are involved and/or the distortion suffered in 'fitting' them into fewer than four dimensions is not too great. However, if the number of points is even moderately large, this fitting process may not be computationally feasible or, the inherent dimensionality of the points may be such that they do not all 'fit' in fewer than four dimensions without suffering an unbearable

amount of distortion. In either event, an expedient would be to find natural 'clusters' of points, if present, whose average within-cluster similarities are high compared to the between-cluster similarities. As a general guide, we want to find the same clusters as those we would pick out if we could actually see the arrangement of the points.

After determining the clusters and some measure of distance between them, hopefully we might then fit the clusters into three or fewer dimensions with a reasonably small amount of distortion. Even if the clusters don't 'fit', knowing their membership allows us to compute their average distances apart as well as statistics (length, principal components, etc.) for each cluster.

2.4 Cluster Analysis

The remainder of this thesis concerns this problem of 'clustering' or 'cluster analysis'. Forgy (1965) makes the distinction between 'cluster analysis' and factor analytic techniques in the following way: "Cluster analysis is any procedure that gives primary attention to relationships among observations (OTU's), or cases, rather than among variables (attributes, characters)". The term 'cluster' above (2.3) is deliberately left undefined since none of the many specific definitions

which have been proposed seem adequate or 'best' in any general sense. Indeed, the scope of numerical classification is so broad that the judgement of the user and the purpose for which the clustering is being created may be the ultimate criteria for evaluating or defining the meaning of the term. That is, it depends upon whether the clustering obtained is to be used to determine the agreement of the data with an a-priori hypothesis about the OTU relations or to be used primarily as a descriptive organization of the data. Some clustering criteria (i.e. definitions of clusters or criteria for admittance of an OTU into a cluster) are surely not sufficiently restrictive to determine the significance of the data; nevertheless, the clustering may provide a description of the data which is adequate to suggest new experiments or new interpretations. It is important, when criticizing or evaluating clustering techniques, to keep this in mind. One of the common faults of proponents of clustering techniques is their failure to distinguish between these objectives. For example, one type or procedure might seek the k-group minimum variance partition of a set of OTU's. Such a partition of the population will always exist, even though the majority of people would agree no

clusters existed in the population. Even if k 'natural' clusters did exist, the minimum variance partition (k-group) need not isolate these clusters. An example taken from Forgy (1965) illustrates this point very well. The figure below (p.16) is a picture of some classical data in the field of astronomy, that of Hertzsprung and Russell, which plots the distribution of stars with respect to temperature and absolute luminosity. The obvious natural grouping is into two very large clusters or classes of stars, the main sequence stars and the red giants (see diagram). However, the 2-group minimum variance partition in this case cuts right across the (intuitively) natural partition.

The detection and description of such natural clusters is a challenging and interesting problem whose solution would have broad applications. For most purposes, a classification based on natural clusters would be the most generally useful. For this reason, we will be primarily concerned with this problem although we will briefly describe a method proposed by Ward (1963) for finding certain minimum variance partitions.

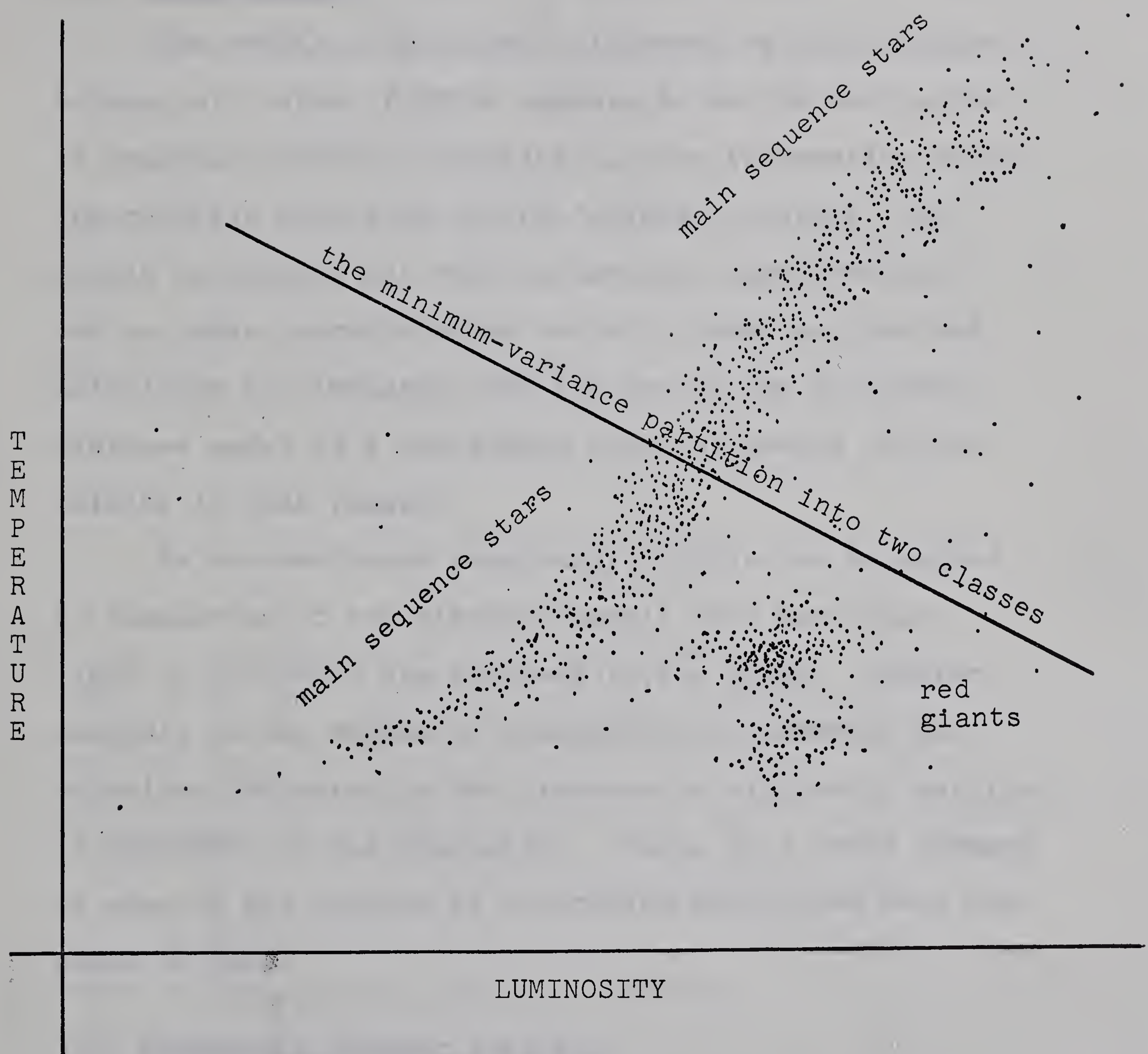
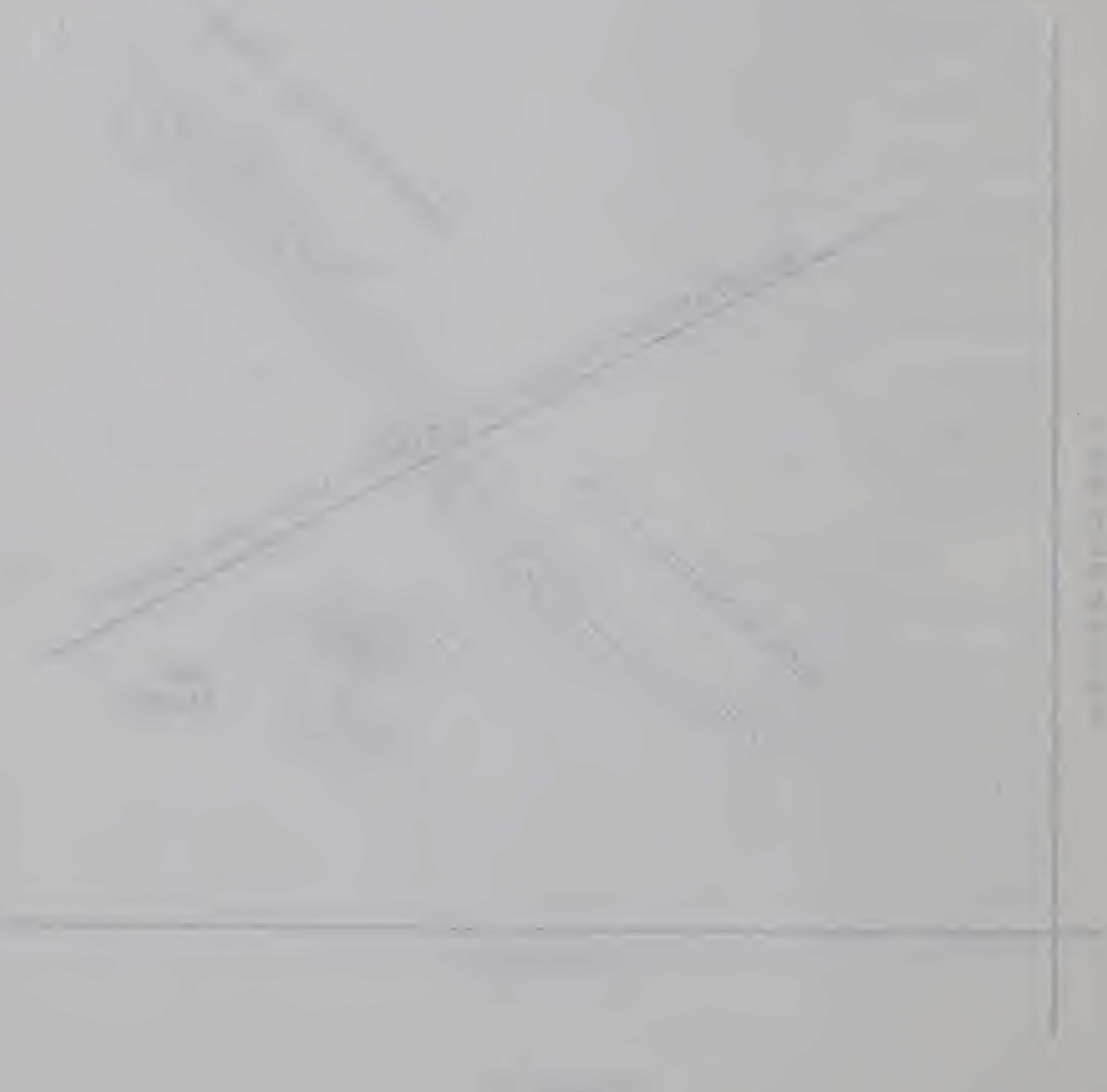

main sequence stars
the minimum-variance partition into two classes
main sequence stars
red giants
TEMPERATURE
LUMINOSITY

Figure 2



CHAPTER III

CLUSTERING PROCEDURES - A BRIEF SURVEY

3.1 Introduction

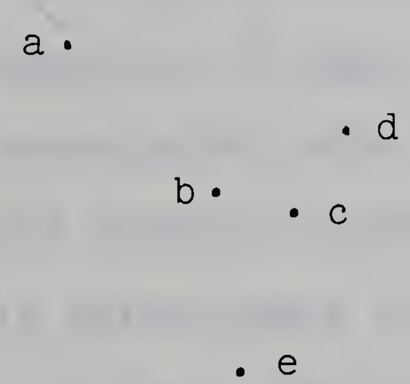
The matrix of Euclidean distances or similarities between all pairs of OTU's appears to be the best point of departure since it contains all the information about the relative positions of the 'points' in space. It should be pointed out that by various transformations one can make such distances reflect almost any desired definition of similarity and the use of the Euclidean distance model as a descriptor does not demand inflexibility in this respect.

As was mentioned previously, little can be gained by inspection of the distance matrix when more than eight or ten OTU's are involved in the study. Cluster analysis is one method of attempting to summarize the relations indicated by the distance or similarity matrices of the OTU's of the population. Below is a brief summary of some of the methods of clustering which have been proposed to date.

3.2 Elementary Cluster Analysis

This method, described by Sokal and Sneath (1963),

is the simplest and most naive approach to the problem of clustering. A level on the scale of similarity coefficients is selected arbitrarily and those pairs of OTU's whose similarity is above that level are written down and the relationships expressed by their coefficients are indicated by lines or links connecting the OTU's, which are represented as points. As the cluster criterion or resolution parameter is lowered, more distant points are combined. Below is an example of a set of points and a dendogram which would result from the application of this method.



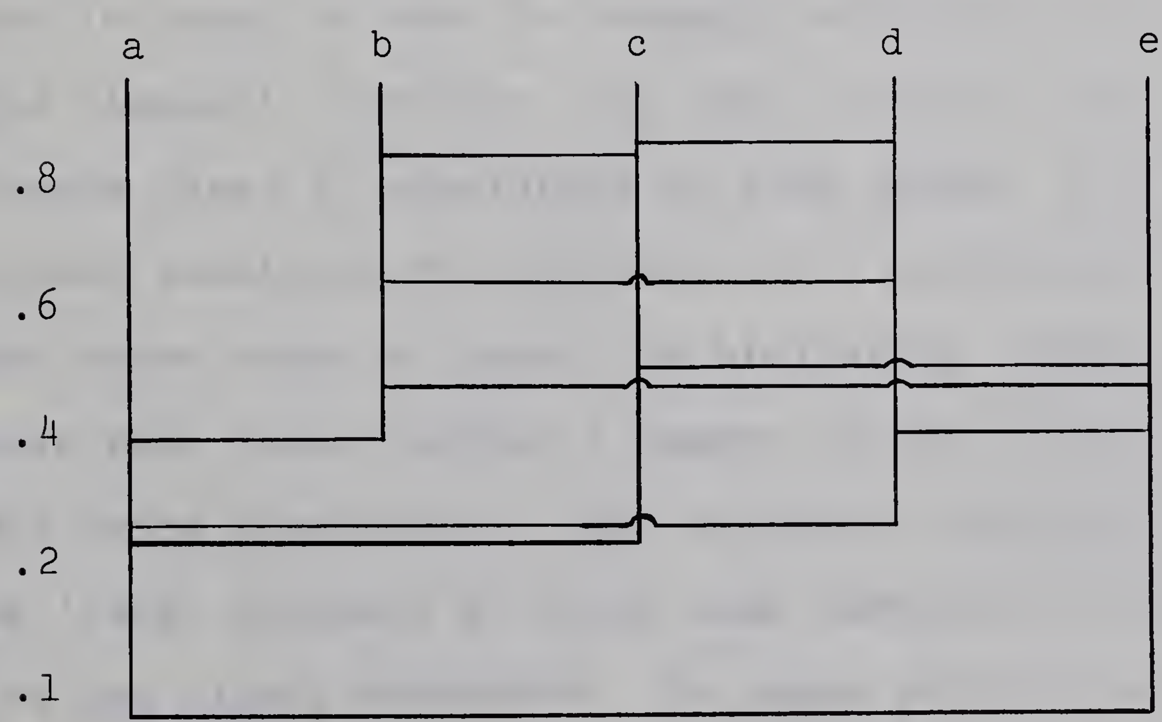


Figure 3

This method is unsatisfactory in most applications since the 'dendogram' produced is not really a reduction in the data because every relation between each pair of points is displayed. For more than a few points, such a dendogram would be unreadable.

3.3 Single Linkage Method

Initially, the resolution parameter for the single linkage method is set relatively high. A cluster is formed using a pair of points whose similarity is above the parameter and others are admitted if they too are above that level. When all the clustering at this level is completed, the level is gradually lowered allowing new members (and clusters) to join established clusters and new clusters to form. Admission of an OTU (or another cluster) to a cluster is based on what is commonly referred to as the 'single linkage' criterion. By this is meant that if the admittance level of similarity is some number 'c', a sufficient condition for admission of a particular OTU is that there exist at least one similarity linkage at or above that level between a member of the cluster and the OTU being considered. This technique obviously allows 'long' clusters in which some members of the cluster are widely separated. In cases such as the one below, this might be desirable.

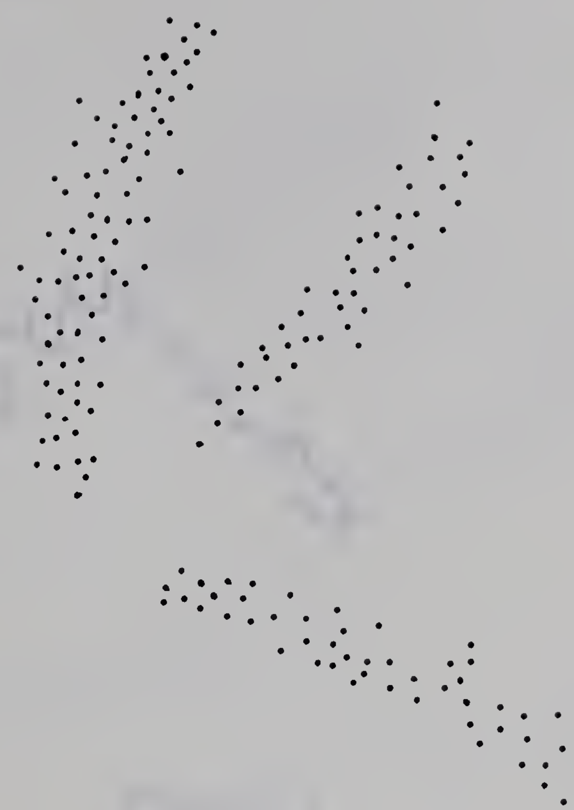


Figure 4

In other cases, the ability to form long chains may lead to unsatisfactory 'clusters', i.e. if noise is present, 'bridging' may occur and the results will be erratic. The unhappy situation below, where the lines join one group of points which have been "clustered" at a particular level, may result.

Clearly some further criterion must be introduced to make the method useful in the majority of practical applications where experimental error and 'wildshots' are likely to be present in the data.

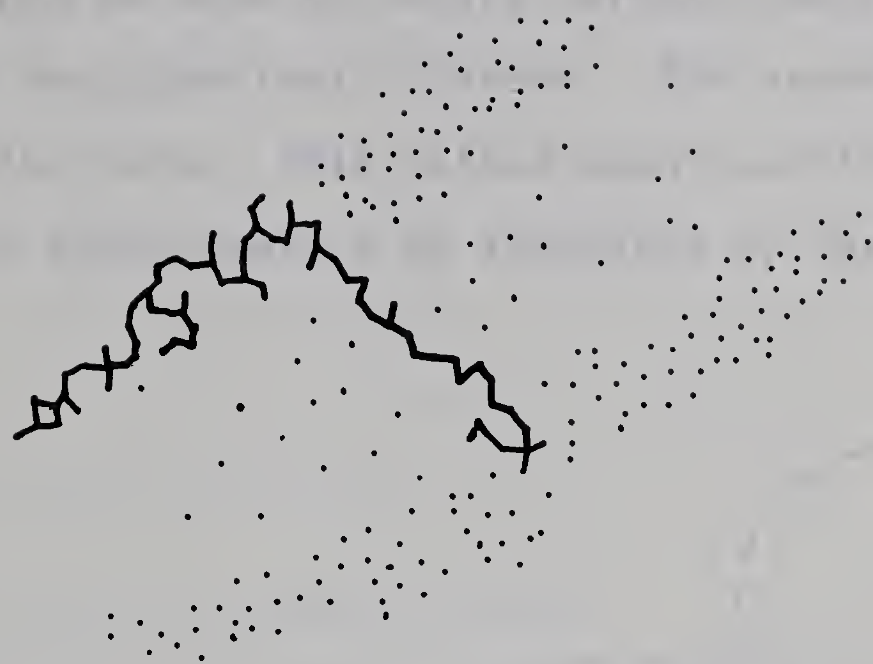


Figure 5

3.4 Clustering by Complete Linkage

The method of complete linkage is similar to the method described above (Single Linkage) except that admission of an OTU to a cluster is made on the basis of the 'complete linkage' criterion. That is, a prospective OTU must have a similarity above a certain level with every member of the cluster in order to gain admission. After all clustering has been done at a certain level, the level is lowered and admissions to each cluster are again attempted. Clusters are not combined until the most widely separated points in the two clusters are above the clustering level. This method of

clustering keeps the clusters spherical and for this reason may not be able to detect certain configurations of elongate and spherical clusters. For example, at one clustering level, this method might partition the points below approximately as indicated by the dotted lines.

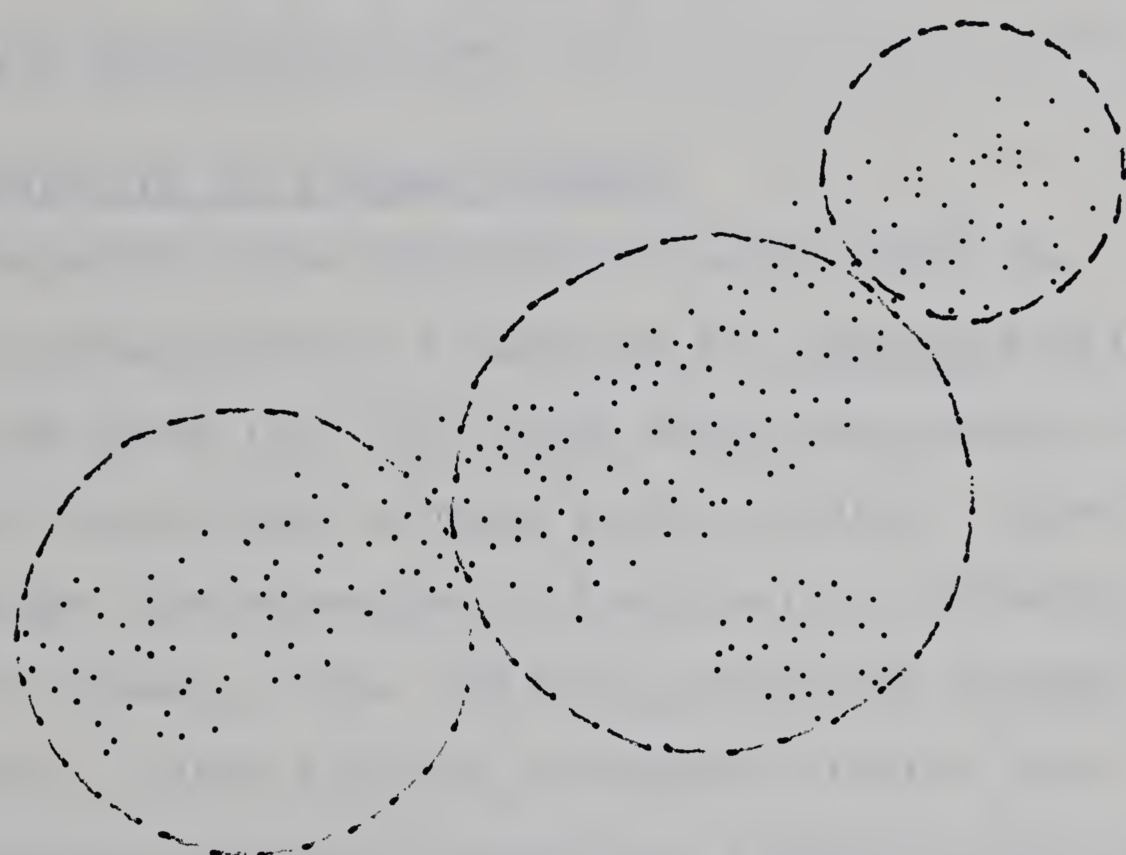


Figure 6

In addition, the final outcome may be strongly influenced by the pair of points used as starting values for each cluster.

A refinement of this method can be made in the following way. Where a prospective member lowers the average similarity of the cluster by more than a prescribed amount, the OTU is not admitted and clustering about that group is terminated. The problem with this practice is that it is difficult to determine what the prescribed amount should be.

3.5 Clustering by Average Linkage

Admission of an OTU into a cluster using the average linkage method is based on the average similarity of the new point (i.e. the point being considered for admission) with those already in the cluster. Except for this change, the procedure is identical to the method of complete linkage. This criterion, involving average similarity, allows slightly elongated clusters since the criterion essentially maeasures distance from the centroid of the cluster rather than the most distant point in the cluster. When all clustering has been done at a particular level, the similarity coefficients are recalculated among clusters already formed at that level as well as between all clusters and those OTU's which have remained isolated. The procedure is then repeated using a lower similarity level. This method

suffers from the defect that it bases admittance on highest average similarity (distance from centroid) rather than the 'best link'. For example, in Figure 10, points 1, 2 and 3 should be excluded, although they are closer to the centroid than some of the members of the cluster.

3.6 Central or Nodal Clustering (The Method of Rogers and Tanimoto)

Given similarities between t OTU's to be classified, the method can be described as follows:

First a value R_i is obtained which symbolizes the number of non-zero similarity coefficients OTU i has with other OTU's. Alternatively, in the case of continuous variables, which incidentally was not Roger's and Tanimoto's case, R_i might symbolize the number of similarity coefficients above a certain level (see Silvestri et al. 1962). These values are indices of typicality or 'degree of relationship'. The greater the R_i value, the more typical of the whole group of OTU's under consideration or more central in a geometric sense is the OTU. Since some of the OTU's may have the same R_i value, a finer, more discriminate measure of centrality is needed. The value

$$H_i = \sum_j(-\log_2 S_{ij}) = \sum_j d_{ij}$$

is computed for each OTU i. Since a small fluctuation in an R_i value may produce a large change in its rank order, a value $T_i = H_i/R_i$ is computed which ranks the OTU's simultaneously on their degree of relationship and on the number of OTU's to which the OTU is related. Note that Rogers and Tanimoto define $d_{ij} = -\log_2 S_{ij}$. This has the effect of magnifying the 'distances' between relatively dissimilar OTU's.

Since the method of clustering depends on a measure of 'inhomogeneity' for each 'provisional' cluster formed, we will briefly outline the considerations upon which the measure is based. The value $\varepsilon_n = \log_2[(n/2)(n-1)]$ is the maximum value that can be attained by an entropy function associated with $(n/2)(n-1)$ segments between objects in n-space where the probability of selecting any particular segment is a constant equal to $1/[(n/2)(n-1)]$. However, taking into account the number of identical OTU's (g), and the number of unrelated OTU's (h), the maximum entropy of the system becomes:

$$\varepsilon_n = \log_2[(\frac{n(n-1)}{2} - g) - h]$$

In a similar way, we define the total entropy of a given set of points determined by the OTU's whose distances are the elements of the matrix (d_{ij}) by:

$$E_n[(d_{ij})] = -1/2 \sum_{i,j}{}' \left(\frac{d_{ij}}{Q} \log_2 \frac{d_{ij}}{Q}\right)$$

$$\text{where} \quad Q = 1/2 \sum_{i,j}{}' d_{ij}$$

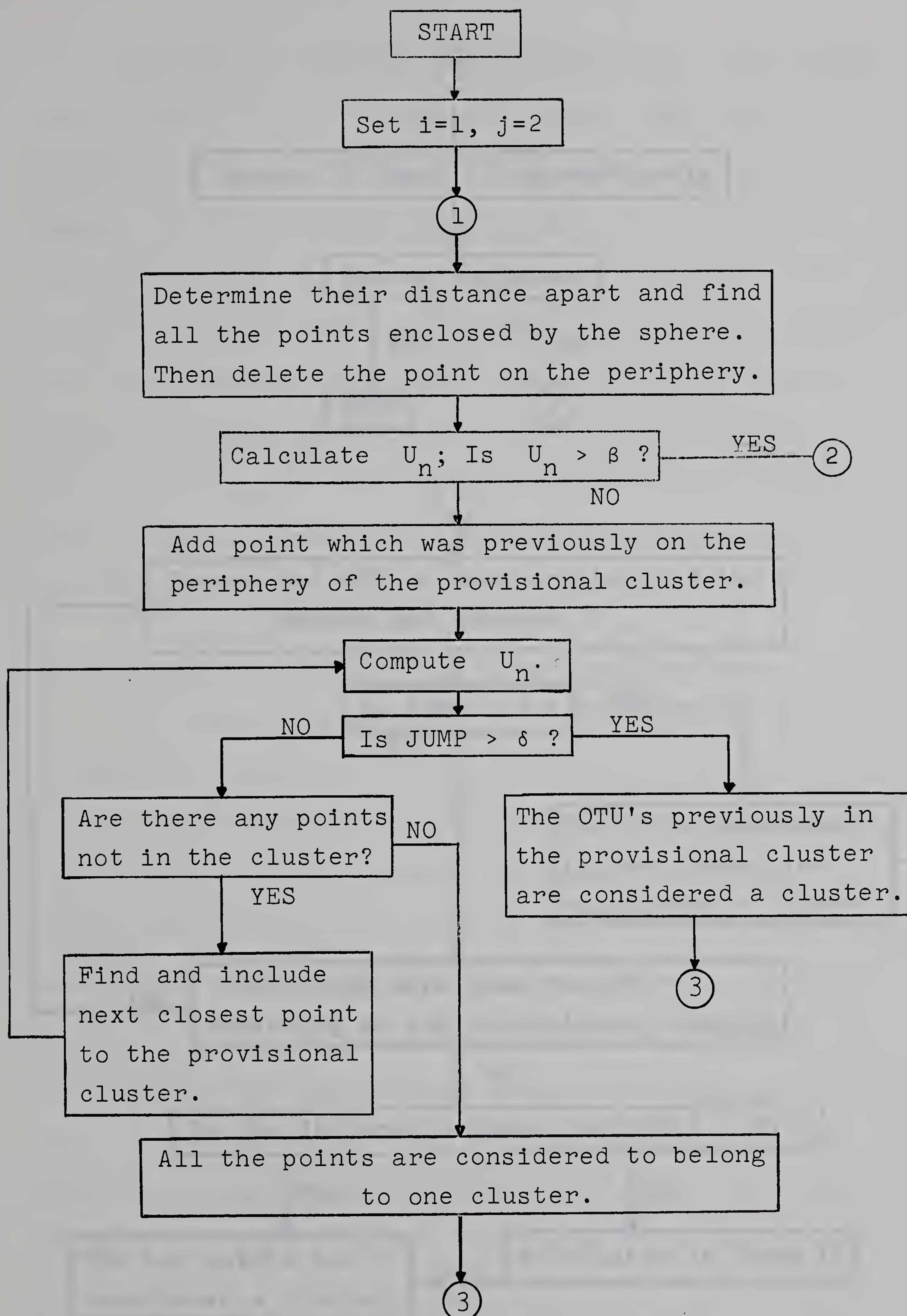
and $\sum'$ indicates summation only over finite elements after repeated rows and columns have been deleted. Our measure of inhomogeneity then becomes:

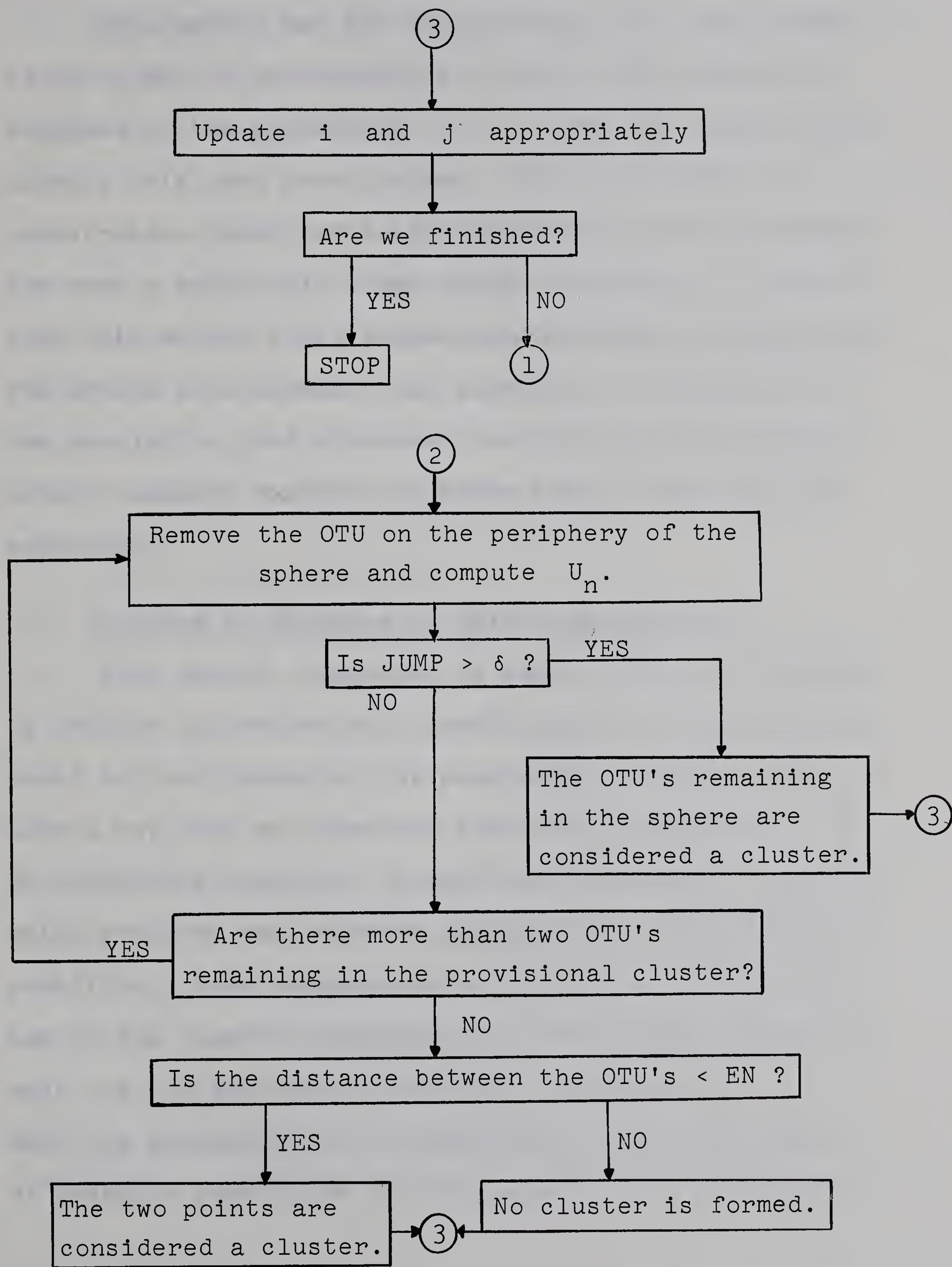
$$U_n[(d_{ij})] = \frac{\varepsilon_n(g,h) - E_n[(d_{ij})]}{\varepsilon_n(g,h)} = 1 - \frac{E_n[(d_{ij})]}{\varepsilon_n(g,h)}$$

The actual clustering process or analysis which Rogers and Tanimoto propose can be summarized in the following way. Three constants come into play in the procedure. The first constant, β, defines the amount of inhomogeneity we consider allowable for the OTU's

currently under study and still group them together. The second, EN , defines the amount of resemblance deemed necessary between two points before they are allowed in the same cluster. The third constant, δ , defines the size of the 'jump' in the inhomogeneity (either positive or negative) required to separate a valid cluster from a preformed, provisional one. U_n is first computed for the whole population and if it is greater than β , the T values of the OTU's are ranked smallest to largest and the procedure outlined by the flowchart below is carried out.

Let i run from 1 through n and for each i let j run from i through n except $j \neq i$. OTU i is to be made the center of the provisional hyper-spherical cluster and OTU j is to be on the surface of the hypersphere.


START
Set i=1, j=2
1
Determine their distance apart and find all the points enclosed by the sphere. Then delete the point on the periphery.
Calculate U_n; Is $U_n > \beta$?
YES
2
NO
Add point which was previously on the periphery of the provisional cluster.
Compute U_n.
Is JUMP > δ ?
NO
YES
Are there any points not in the cluster?
NO
YES
The OTU's previously in the provisional cluster are considered a cluster.
3
Find and include next closest point to the provisional cluster.
All the points are considered to belong to one cluster.
3


3
Update i and j appropriately
Are we finished?
YES
NO
STOP
1
2
Remove the OTU on the periphery of the sphere and compute U_n.
Is JUMP > δ ?
YES
NO
The OTU's remaining in the sphere are considered a cluster.
3
Are there more than two OTU's remaining in the provisional cluster?
YES
NO
Is the distance between the OTU's < EN ?
YES
NO
The two points are considered a cluster.
3
No cluster is formed.

This method has the disadvantage that there seems to be no way of determining the values that should be assigned to the parameters β , δ , and EN , apart from using a trial and error scheme. This introduces an undesirable element of subjectiveness into the procedure. For even a moderately large number of points, it appears that this method would become computationally impractical. The method also assumes that clusters, if they exist in the population, are spherical, and it will not reliably detect elongate parallel clusters even if they are well separated.

3.7 Grouping to Optimize an Objective Function

This method, suggested by Ward (1963), is designed to produce a hierarchical classification so that, at any level in the hierarchy, the population is partitioned in such a way that an objective function is optimized. By an 'objective function' is meant any functional relation which reflects the relative desirability of a particular partition. Ward suggests as an objective function the sum of the cluster variances (a cluster might consist of only one OTU and would then have a variance of 0.0). When the population is of appreciable size, the number of possible partitions of the population is prodigious

and investigation of every one is a practical impossibility.[2] Ward proposes an alternative: Beginning with n groups (of one point each), we compute the value of the objective function for the $n(n-1)/2$ possible $(n-1)$ group partitions which can be created by combining a pair of groups taken from the original n groups. The combination is made which minimizes the increase (decrease) in the objective function being minimized (maximized), resulting in $n-1$ groups. The process is repeated beginning with these $n-1$ groups until the number of groups is reduced to one.

3.8 ISODATA (Ball and Hall)

ISODATA is an iterative procedure for reducing a set of OTU's represented by points to a set of 'average points', each one representing a set of points whose variance about the average point is small. A set of points is chosen initially (in some way) and by constantly changing the members of each group, the method systematically reduces the variance about the average points through combination and 'splitting' of the points according to user-supplied parameters which set limits on

---

[2] The number of possible partitions of 10 objects is 123,577.

the amount of variance and the number of points allowed in any cluster. After 'lumping' and/or 'splitting', the modified set of average points is used as the starting points for the next iteration.

The proponents of the method claim that the resulting points obtained from the procedure are an 'adequate' description of the data. They do not, however, make clear for what use the points are 'adequate'. The average points obtained do not necessarily represent 'natural clusters', and the method can not, therefore, be classed as a cluster-seeking method.

3.9 Factor Analysis

Factor analysis, when used in the classification process, is generally used to gain information about the over-all properties of the OTU population. For example, a principle components analysis of the correlations between characters measured may be carried out to investigate how several measures of the OTU's might be combined to produce maximum descrimination among OTU's along a single dimension. A similar analysis of correlations between OTU's may reveal that several independent dimensions are necessary to define adequately the domain or space under study. Factor analysis may then

be used to reduce the dimensionality of a set of variables (OTU's) by taking advantage of these inter-correlations. This is done by defining the OTU's in a space whose coordinate axes are defined using only the principle components which account for significant portions of the over-all variance.

The methods which actually use factor analytic techniques to partition the population vary somewhat as to their details but the essence of them is the following: The projection of the points or OTU's onto the maximum principle component is examined and the population is divided (if possible) on the basis of this 'marginal density'. This is essentially a 'one dimensional view' of the OTU's. If all, or nearly all, of the variation is expressed on the first component axis, the other axes may be ignored. If not, the divided population may be subdivided again on the basis of the marginal density along the second component axis and so on until the variance along the axes becomes too small to be worthwhile. The problem with this scheme is that cases can arise where the clusters are arranged in the space so that no variation in the marginal density is obvious along any of the principle axes. This is the case in the example below.

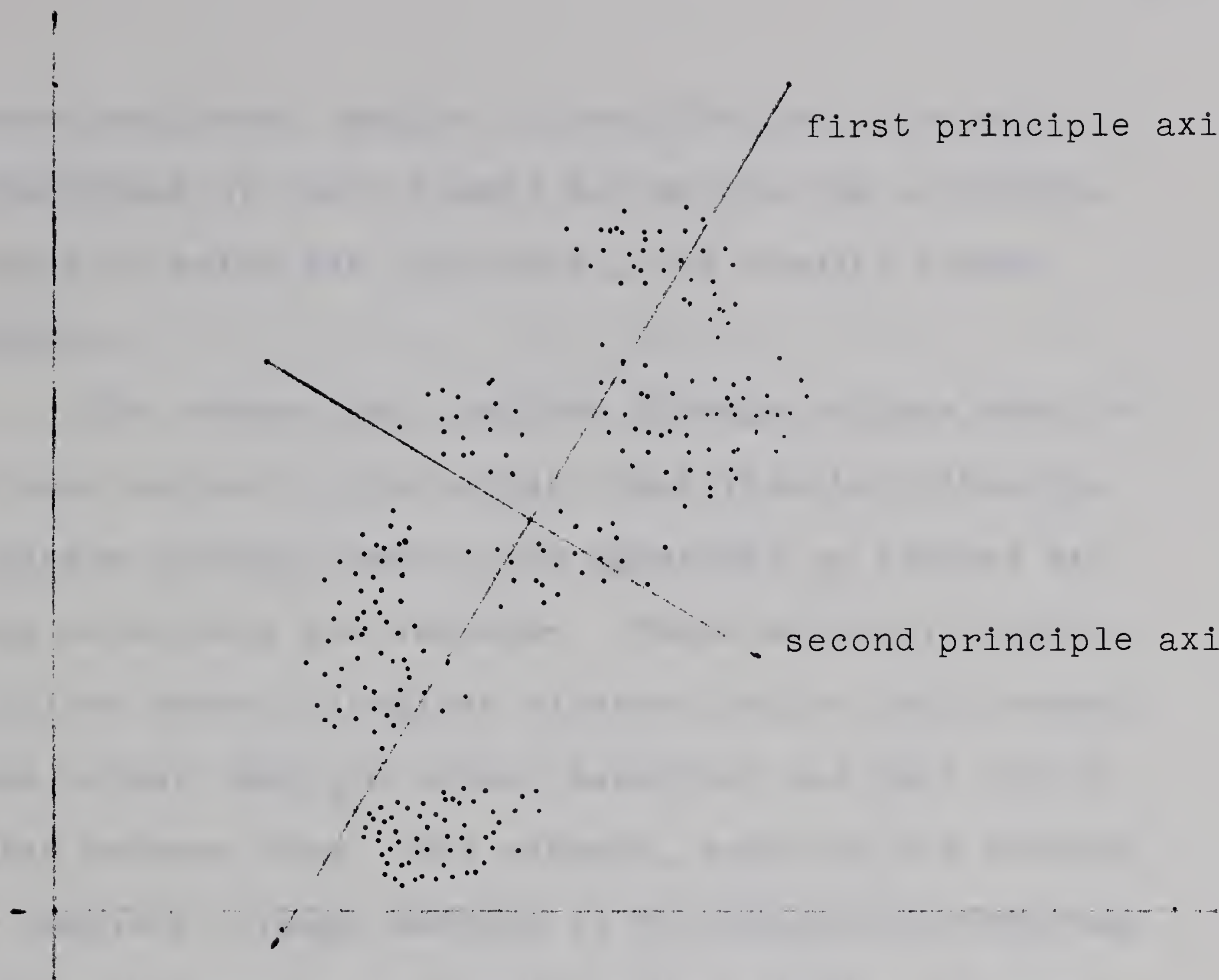
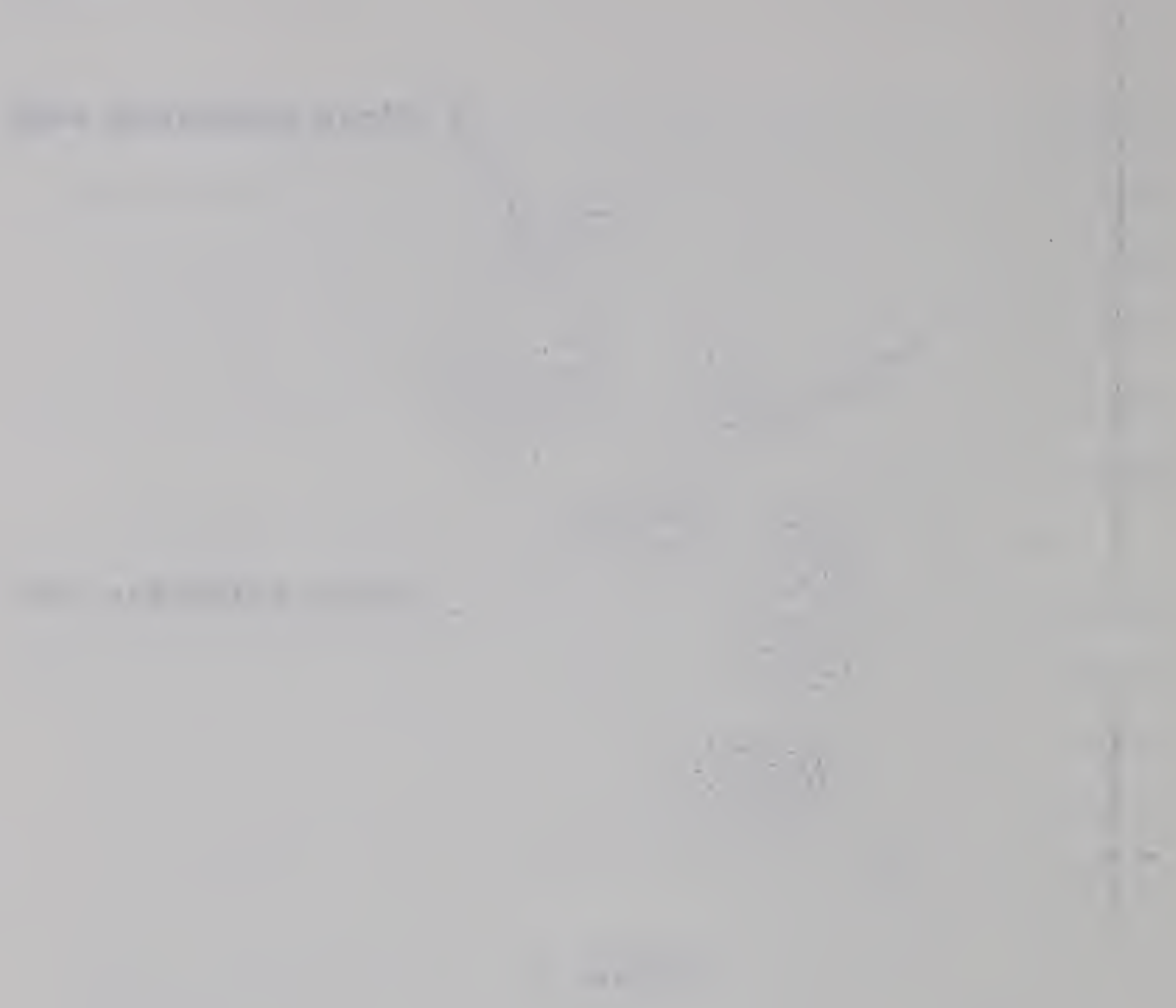


Figure 7

3.10 Comparison and Evaluation of the Methods

As was mentioned before, the single linkage criterion method allows long straight or serpentine rows where each point may be linked to only one or two others and where the similarity between end points in the 'cluster' may be very low. Such linked strings, in biological taxonomy, are likely to be artifacts, but may be of interest in other applications. Forgy (1964) found that this method



worked well when applied to very distinct clusters (regardless of their shape) but as soon as a moderate amount of noise was introduced, the results became erratic.

The average and complete linkage methods seem to be more suited to biological classification since the clusters produced tend to be spherical or globose and have relatively low variance. These methods, however, will not detect elongated clusters and/or oddly shaped ones unless they are widely separated and have little noise between them. For example, applying the average or complete linkage methods to the Russell-Hertzsprung data would produce a classification similar, at some level, to the two-group minimum variance partition.

One problem associated with single, average, and complete linkage methods is that the results are usually expressed as a dendogram for interpretation, and, in some cases, this dendogram can be very misleading. This is due to the many possible equivalent dendograms (formed by rotating the 'branches' as indicated by the arrows in Figure 8) that can be drawn using the output from one of these programs and we are given no indication of which one is the best. Once a particular OTU has been linked

to a cluster, it remains associated with it throughout the lower levels of the classification despite the possibility that it may be virtually midway between that cluster and another in a different branch of the dendogram. For example, the dendograms drawn below for the set of points (Figure 8a) give the misleading implication that A is closer to D than B.

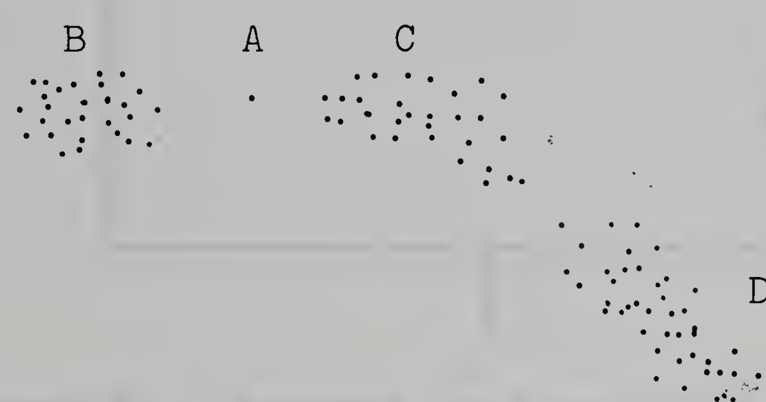


Figure 8a

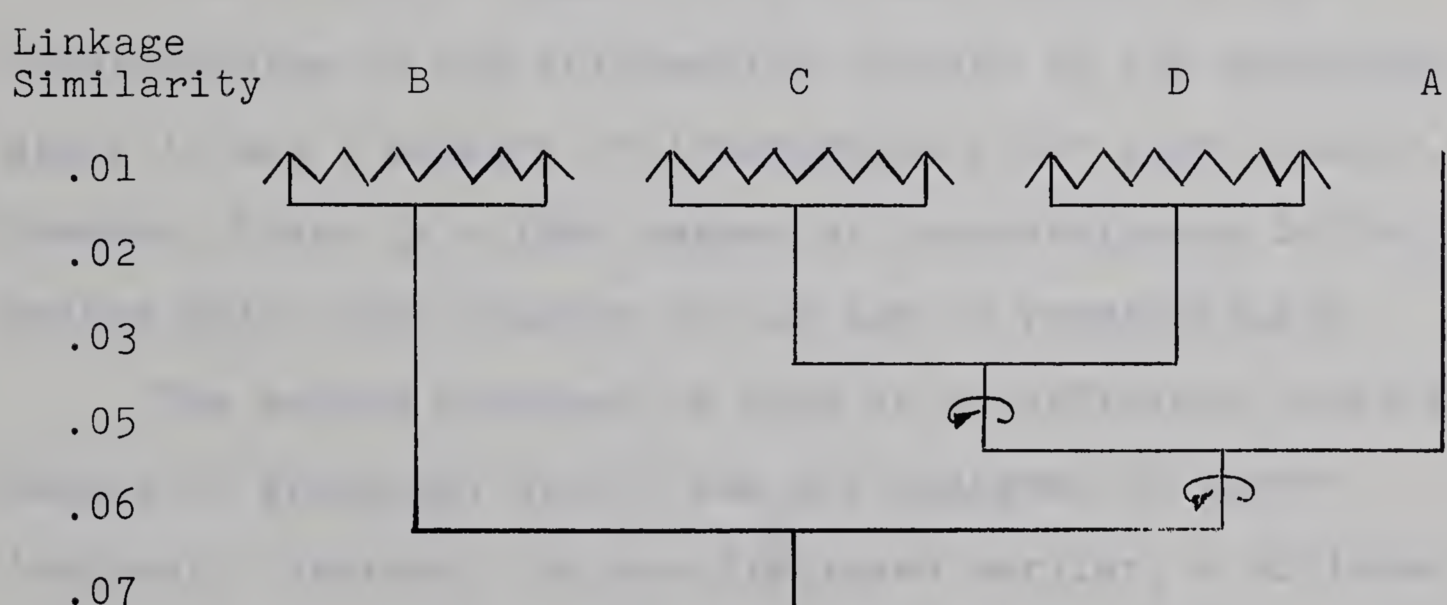


Figure 8b - Single Linkage Dendogram

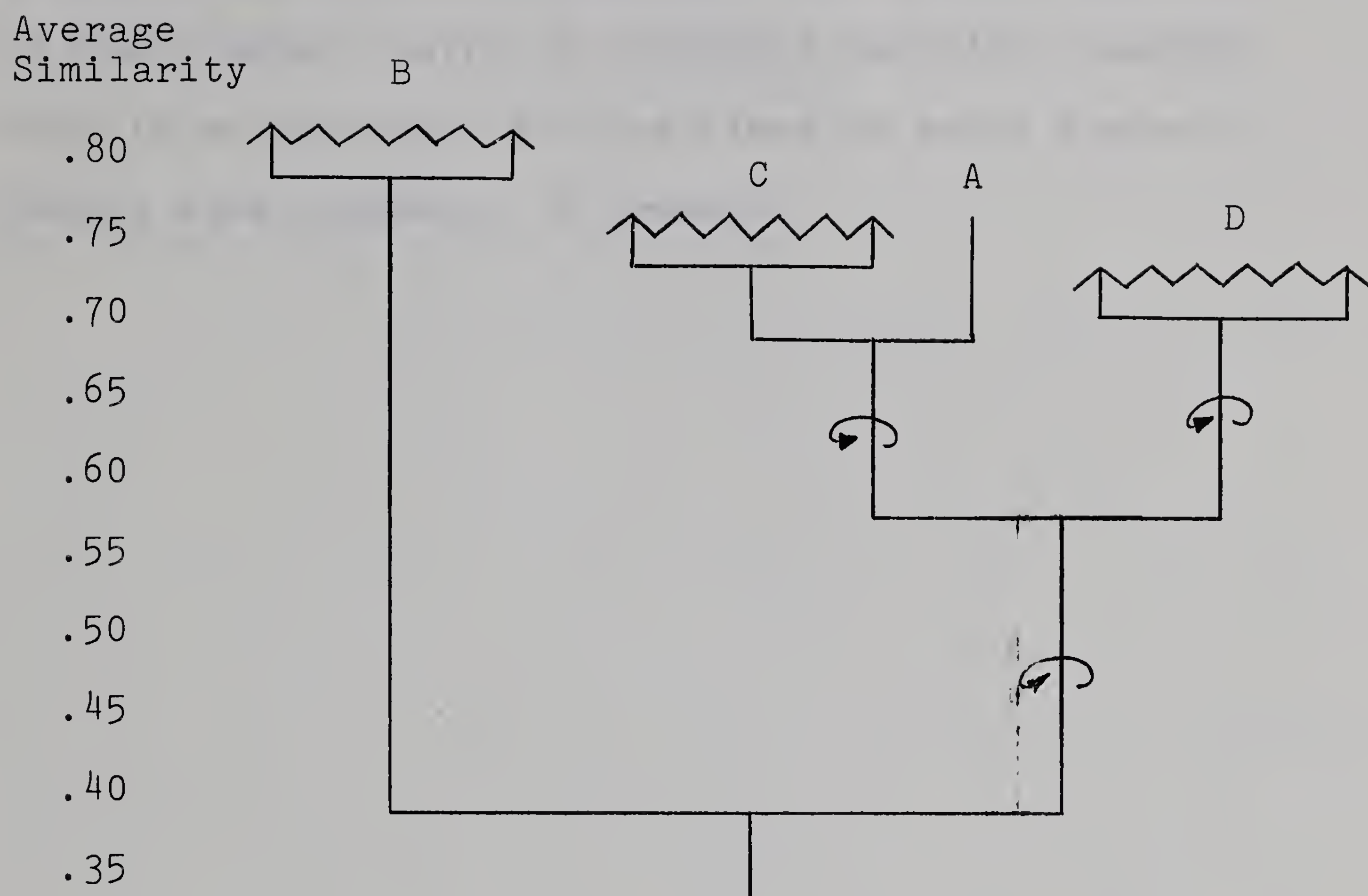


Figure 8c - Average or Complete Linkage Dendogram

The method of Rogers and Tanimoto gives us some understanding of the information content of the groupings since it has a measure of inhomogeneity for each cluster. However, there is a fair degree of indeterminancy in the method which runs counter to our aim of repeatability.

The method proposed by Ward is an efficient objective method of grouping, but it was not designed to detect 'natural' clusters. As was displayed earlier, a minimum-variance partition of the population is not necessarily

a useful one for most purposes. The method will not reliably detect pairs of elongated parallel clusters. This is an important failure since we would prefer to detect such clusters, if present.

CHAPTER IV

A NEW PROCEDURE

4.1 Development of the Procedure

It seems that clustering procedures using only one criterion are either so restrictive that they cannot 'detect' elongate clusters or are so 'permissive' that they become erratic in the presence of noise. Using the average linkage procedure, the point nearest the centroid of the cluster is the first to be considered for admission. As was pointed out in section 3.5, this may not always be the most satisfying choice, since the point outside the cluster which has the highest average similarity with those points in the cluster may still be relatively far from any of them. Using the complete linkage criterion, the points which can be enclosed by a sphere with diameter equal to some function of the resolution parameter, and which encloses the points previously placed in the cluster, are eligible for admission. However, the order in which the points are admitted will determine which of those eligible points will actually be admitted, since the admission of one point may restrict the 'movement' of this imaginary sphere so as to exclude another equally eligible

point. Hence, there is considerable indeterminacy in the method.

Some reflection and considerable experiment yielded the following ideas. If there are any clusters in the population, then surely the closest pair of points will be members of one of them. Do these two points form a cluster by themselves, or are there other members? The next candidate for admission to the cluster should logically be the point closest to either of the initial two. From this comes the reasonable extension that the point which should be considered next for admission to any cluster is that one which is closest to a point already in the cluster.

This search for additional members was made more efficient by replacing the traditional similarity matrix with a vector of ranked similarities and their corresponding pairs of points.

Given this basis for the order of selecting points for possible admission to the cluster, the question then arose concerning the basis upon which we should terminate additions to the cluster. The answer which came to mind was to terminate additions to the cluster if the prospective point was 'much' farther away than the last point admitted,

i.e. if there was a discontinuity in 'closeness'. A measure of closeness to all the members of the cluster is the average linkage (similarity) with them. Therefore, a sudden drop in average linkage indicates a discontinuity in 'closeness', that is, there is a relatively large space around the members already in the cluster. Any drop in average linkage can be accentuated by subtracting that drop from the new average linkage. That is, by computing:

DIFF = (NEW AVERAGE - DROP)

where

DROP = (OLD AVERAGE - NEW AVERAGE)

A way to look for discontinuities is to set a lower limit on the value DIFF. Since we do not know, in advance, what this lower limit should be, and since clusters of various sizes and shapes may be present, it is necessary to repeat the procedure using different values; that is, to examine the population at different

levels of resolution. This procedure alone is sufficient to detect more or less 'globose' clusters. However, problems arose when points which were near the centroid of an elongate cluster, but which were still rather far from any point in the cluster, became eligible for admission. To prevent such admissions until a lower level of resolution, the addition to the procedure of some type of single linkage criterion was necessary. Unlike the average linkage, the size of the single linkage of successive prospective members may vary erratically. Therefore, the new single link was compared to the average of the preceding single links rather than just the previous one. This is reasonable because the average of the single links, like the average linkage of a single point, is determined by the configuration of all the points in the cluster. Again, since we wanted to accentuate any discontinuity, we computed the following quantities and set a lower limit on the value JUMP:

JUMP = (2 x NEWLINK - (AVERAGE OF PRECEDING LINKS))

which is equivalent to

(NEWLINK - (AVERAGE OF PRECEDING LINKS - NEWLINK))

An example where this criterion is applicable appears below (Figure 9). Although point 1 has a higher average similarity with the cluster points than does point 2, it should not (intuitively) be admitted until a low level of resolution.

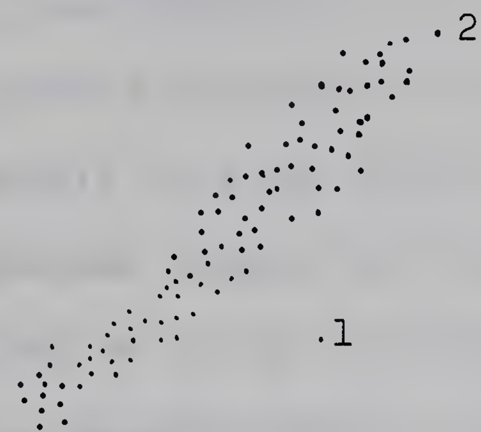


Figure 9

For some configurations, another criterion was found necessary to deal with the same general type of occurrence at lower levels of resolution. This criterion, which we call the 'ratio criterion', was based on the following rationale: if the rank of the similarities, during the growth of a cluster, dictates that the cluster become large in some dimension(s) before considering points

for admission which are near the centroid of the cluster in other dimensions, then intuitively these points should be denied admittance until a low level of resolution. This was effected by setting a lower limit on the ratio (MINSIM/FARSIM) where FARSIM is the minimum similarity between the point being considered for admission and any point in the cluster, and MINSIM is the minimum similarity between any pair of points already in the cluster. This ratio cannot become small at high levels of resolution, but as the clusters become large, and the similarities range down to the order of .5 or .4, this restraint becomes powerful. Below is an example which illustrates the application of this criterion. The maximum distance of point number 4 from any member of the cluster is much greater than the maximum distance of points 1, 2, and 3 from any other member of the cluster. Points 1, 2, and 3 would not be admitted until a relatively low level of similarity.

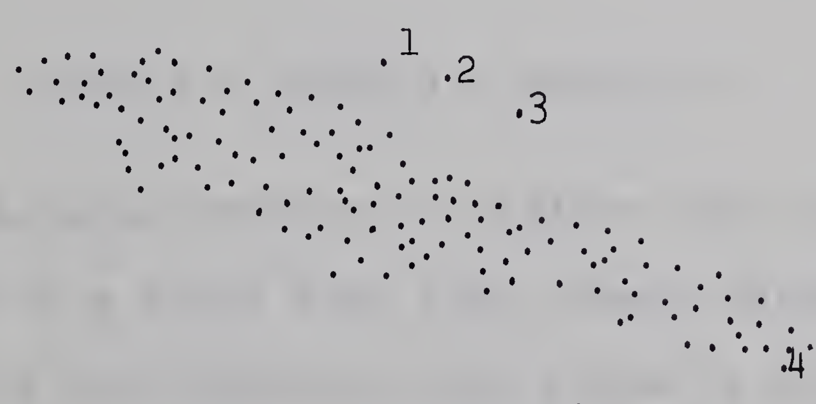


Figure 10

Since our objective was to find natural clusters, which implies mutually exclusive classes, our final criterion is that clustering should be terminated if the prospective point has already been admitted to another cluster.

4.2 Description of the Algorithm

The first three criteria outlined in the above section are dependent upon two (related) resolution parameters whose magnitude reflects how closely one is 'looking' at the population. The primary resolution parameter, RESOL1, is systematically lowered for each succeeding resolution level by equal steps starting at a value BEGIN. The secondary resolution parameter, RESOL2, starts at the same value BEGIN but is reduced by only half as large a step at each lower level of resolution. Hence, these values are related in the following way:

$$RESOL2 = (RESOL1 + BEGIN)/2$$

In the program written to utilize this algorithm, BEGIN, is set to a value such that ninety percent of the similarities of the population are below it and RESOL1 is

reduced by twelve equal steps to a value called FINISH which is set so that ninety-five percent of the similarities are above it. Alternate values of the above constants and the number of resolution levels to be examined are optional to the user of the program.

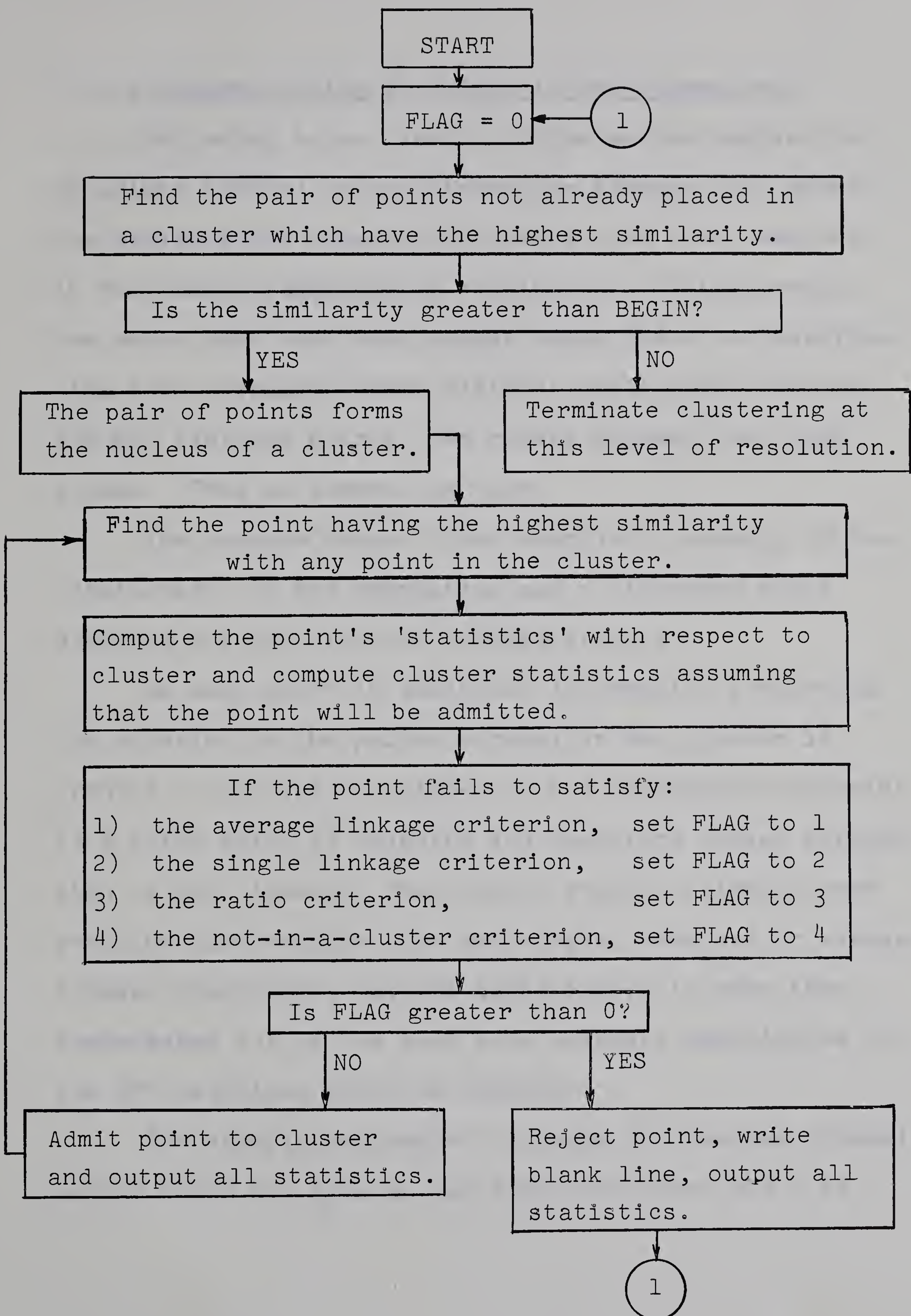
Below are explicit definitions of the quantitative criteria which terminate the additions of points to a cluster.

Terminate additions to a cluster if:

1) DIFF $\leq$ RESOL1 (average linkage criterion)

2) JUMP $\leq$ RESOL2 (single linkage criterion)

3) RATIO $\leq$ RESOL2 (ratio criterion)


START
FLAG = 0
1
Find the pair of points not already placed in a cluster which have the highest similarity.
Is the similarity greater than BEGIN?
YES
NO
The pair of points forms the nucleus of a cluster.
Terminate clustering at this level of resolution.
Find the point having the highest similarity with any point in the cluster.
Compute the point's 'statistics' with respect to cluster and compute cluster statistics assuming that the point will be admitted.
If the point fails to satisfy:
1) the average linkage criterion, set FLAG to 1
2) the single linkage criterion, set FLAG to 2
3) the ratio criterion, set FLAG to 3
4) the not-in-a-cluster criterion, set FLAG to 4
Is FLAG greater than 0?
NO
YES
Admit point to cluster and output all statistics.
Reject point, write blank line, output all statistics.
1

4.3 An Example Using 20 Points in Two Dimensions

Following is an example of the method applied to 20 points (OTU's) in two dimensions (Figure 11), where the similarities between the points have been computed in the manner suggested in section 2.2. Intuitively, one would hope that the process would yield a classification that displayed three distinct multi-point clusters and two isolated points, one midway between two large groups. This is indeed the case.

The program output first supplied a summary of the 'statistics' of the population and a histogram which displays the distribution of similarities.

As each point is admitted, information concerning its relation to the points already in the cluster is printed. One line is skipped if the information pertains to a point which is rejected and therefore caused termination of the cluster. The results require slightly more scrutiny than results from the single, complete or average linkage procedures, but the added effort is more than compensated for by the much more adequate description of the OTU relations which is obtained.

For example, a complete linkage (or average linage) method would not give us any indication that OTU 2 is

midway between two large clusters and does not really belong to either one. Also, an average linkage procedure would place OTU 4 with the lower left-hand cluster before combining any of the large clusters.

Two diagrams of each configuration of points used as examples are included because the labels and dashed lines tend to obscure the relative positions of the points.

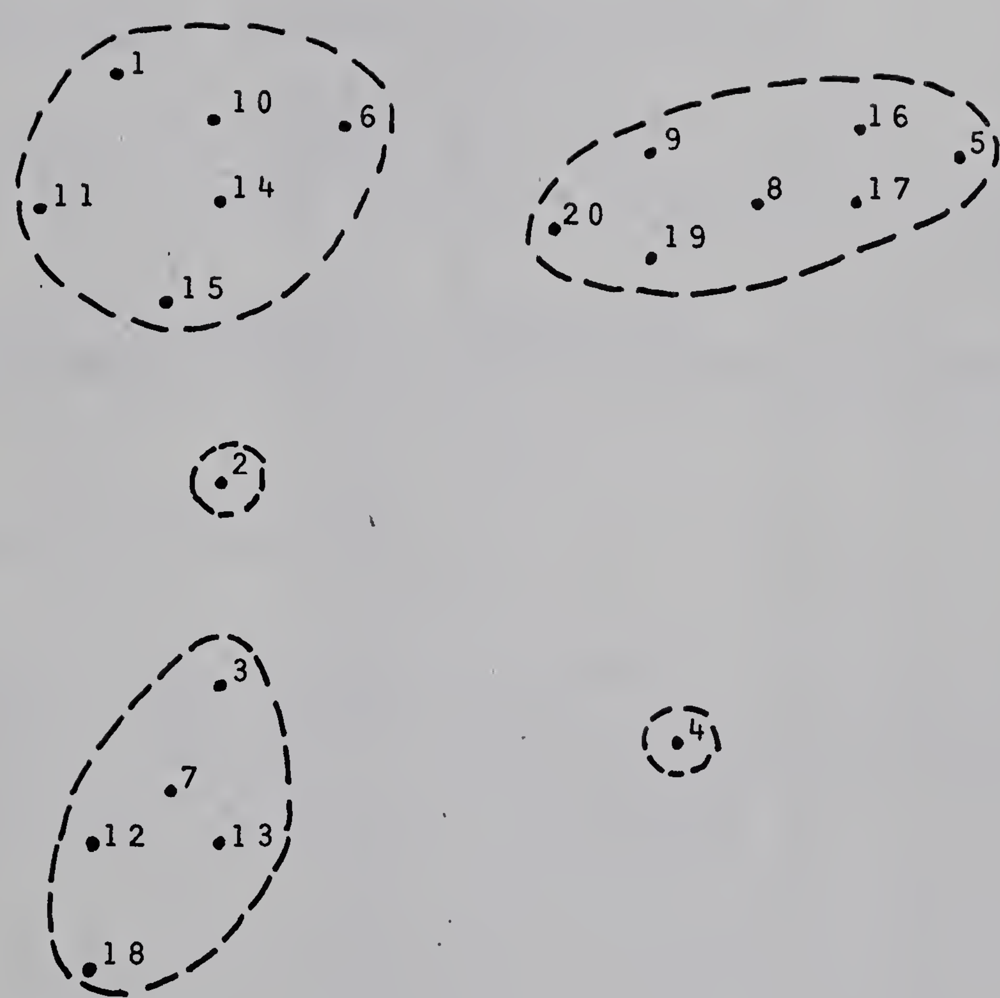


Figure 11

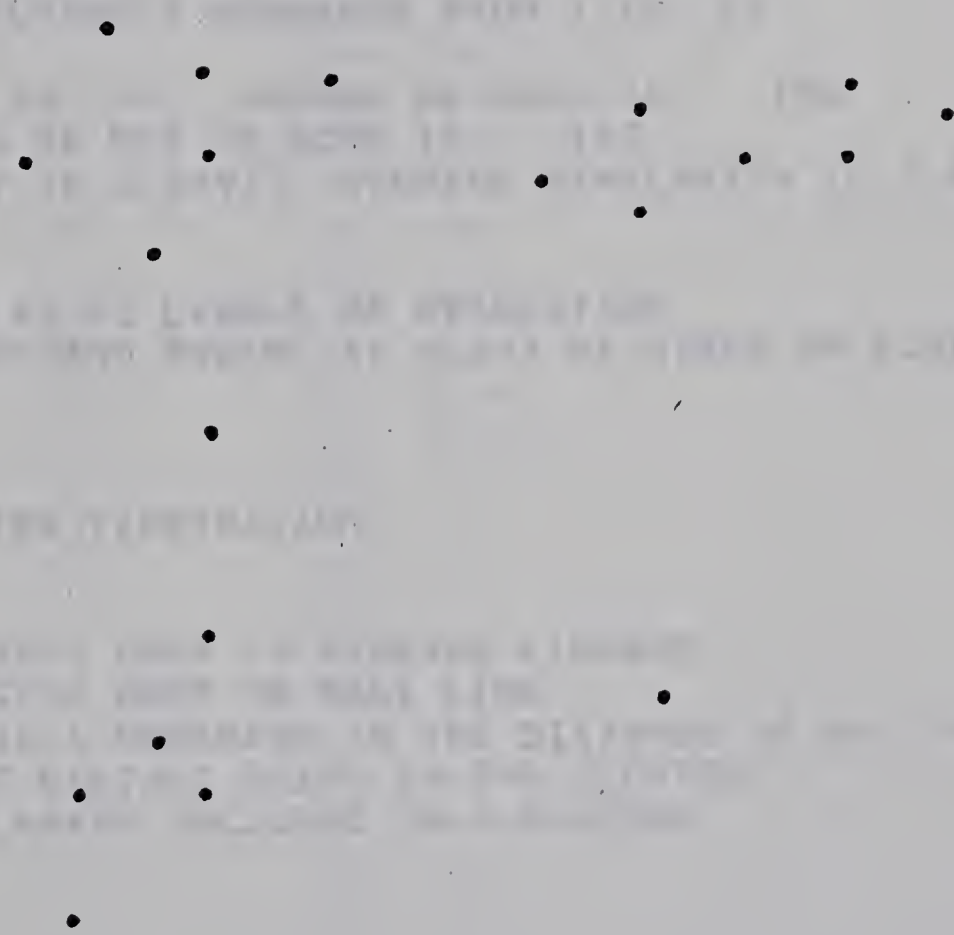


Figure 12

CLUSTER ANALYSIS OF RELATIVE SIMILARITIES

SAMPLE OF 20 POINTS IN TWO-SPACE

POINTS ARE CONSECUTIVELY NUMBERED FROM 1 TO 20

NUMBER OF POINTS IS 20 NUMBER OF SIMS IS 190
NUMBER OF SIMS TO BE PUT IN CORE IS 190
MEDIAN SIMILARITY IS 0.55913 AVERAGE SIMILARITY IS 0.57746

CLUSTERING TO BE AT 12 LEVELS OF RESOLUTION
BEGINNING AT 0.920 AND ENDING AT 0.297 BY STEPS OF 0.057

REASONS FOR CLUSTER TERMINATION

FLAG
COL.1=1 TOO LARGE A DROP IN AVERAGE LINKAGE
COL.2=2 TOO LARGE A DROP IN BEST LINK
COL.3=3 TOO LARGE A DECREASE IN THE DISTANCE OF NEW POINT
FROM MOST DISTANT POINT IN THE CLUSTER
COL.4=4 POINT ALREADY INCLUDED IN A CLUSTER

FREQUENCIES OF SIMILARITIES AT 0.01 LEVELS

LEVEL	NUMBER	TOTAL,	REMAINDER	HISTOGRAM
1.00	0	0	190	
0.99	0	0	190	
0.98	0	0	190	
0.97	0	0	190	
0.96	0	0	190	
0.95	3	3	187	***
0.94	0	3	187	
0.93	3	6	184	***
0.92	8	14	176	********
0.91	5	19	171	*****
0.90	0	19	171	
0.89	3	22	168	***
0.88	2	24	166	**
0.87	1	25	165	*
0.86	3	28	162	***
0.85	4	32	158	****
0.84	5	37	153	*****
0.83	1	38	152	*
0.82	3	41	149	***
0.81	1	42	148	*
0.80	0	42	148	
0.79	1	43	147	*
0.78	0	43	147	
0.77	3	46	144	***
0.76	3	49	141	***
0.75	5	54	136	*****
0.74	0	54	136	
0.73	3	57	133	***
0.72	0	57	133	
0.71	2	59	131	**
0.70	2	61	129	**
0.69	0	61	129	
0.68	4	65	125	****
0.67	1	66	124	*
0.66	2	68	122	**
0.65	2	70	120	**
0.64	2	72	118	**
0.63	3	75	115	***
0.62	2	77	113	**
0.61	4	81	109	****
0.60	3	84	106	***
0.59	3	87	103	***
0.58	6	93	97	******
0.57	1	94	96	*
0.56	4	98	92	****
0.55	2	100	90	**
0.54	3	103	87	***
0.53	5	108	82	*****
0.52	1	109	81	*
0.51	6	115	75	******

0.50	5	120	70	*****
0.49	2	122	68	**
0.48	2	124	66	**
0.47	3	127	63	***
0.46	4	131	59	****
0.45	5	136	54	*****
0.44	5	141	49	*****
0.43	4	145	45	****
0.42	1	146	44	*
0.41	5	151	39	*****
0.40	2	153	37	**
0.39	0	153	37	
0.38	3	156	34	***
0.37	4	160	30	****
0.36	2	162	28	**
0.35	3	165	25	***
0.34	2	167	23	**
0.33	2	169	21	**
0.32	1	170	20	*
0.31	2	172	18	**
0.30	3	175	15	***
0.29	3	178	12	***
0.28	0	178	12	
0.27	0	178	12	
0.26	2	180	10	**
0.25	0	180	10	
0.24	0	180	10	
0.23	2	182	8	**
0.22	1	183	7	*
0.21	2	185	5	**
0.20	0	185	5	
0.19	1	186	4	*
0.18	0	186	4	
0.17	0	186	4	
0.16	1	187	3	*
0.15	0	187	3	
0.14	1	188	2	*
0.13	0	188	2	
0.12	1	189	1	*
0.11	0	189	1	
0.10	0	189	1	
0.09	0	189	1	
0.08	1	190	-0	*
0.07	0	190	-0	
0.06	0	190	-0	
0.05	0	190	-0	
0.04	0	190	-0	
0.03	0	190	-0	
0.02	0	190	-0	
0.01	0	190	-0	

CLUSTERS AT RESOLUTION LEVEL 1

RESOL1 IS 0.920
RESOL2 IS 0.920

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	1200
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	1200
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	1200
4	9										
	19	0.9203									
	20	0.9129	19	0.92	0.909	0.011	0.8980	9	0.90	16	1200

POINTS NOT PLACED IN CLUSTERS
1 2 3 4 5 6 8 11 12 15 18 20

CLUSTERS AT RESOLUTION LEVEL 2

RESOL1 IS 0.863
RESOL2 IS 0.892

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	1200
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	1200
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	1000
4	9										
	19	0.9203									
	20	0.9129	19	0.92	0.909	0.011	0.8980	9	0.90	16	
	8	0.9000	19	0.91	0.887	0.022	0.8650	20	0.84	33	4

POINTS NOT PLACED IN CLUSTERS
1 2 3 4 6 11 15 18

CLUSTERS AT RESOLUTION LEVEL 3

RESOL1 IS 0.807
RESOL2 IS 0.863

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	1000
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	
	1	0.8843	10	0.91	0.867	0.016	0.8502	15	0.82	39	
	6	0.8709	10	0.90	0.851	0.016	0.8352	15	0.80	42	
	11	0.8619	1	0.88	0.844	0.007	0.8367	6	0.75	49	200
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	1000
4	9										
	19	0.9203									
	20	0.9129	19	0.92	0.909	0.011	0.8980	9	0.90	16	
	8	0.9000	19	0.91	0.887	0.022	0.8650	20	0.84	33	4

POINTS NOT PLACED IN CLUSTERS
2 4 11 18

CLUSTERS AT RESOLUTION LEVEL 4

RESOL1 IS 0.750
RESOL2 IS 0.835

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU- BEST LINK	SIM- BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	1200
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	
	1	0.8843	10	0.91	0.867	0.016	0.8502	15	0.82	39	
	6	0.8709	10	0.90	0.851	0.016	0.8352	15	0.80	42	
	11	0.8619	1	0.88	0.844	0.007	0.8367	6	0.75	49	
	2	0.8286	15	0.86	0.745	0.098	0.6472	1	0.67	65	1200
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	
	9	0.8667	19	0.92	0.853	-0.027	0.8803	5	0.76	44	
	20	0.8496	19	0.92	0.807	0.047	0.7604	5	0.68	62	
	6	0.8058	20	0.82	0.674	0.133	0.5414	5	0.52	105	1204

POINTS NOT PLACED IN CLUSTERS

2 4

CLUSTERS AT RESOLUTION LEVEL 5 RESOL1 IS 0.693
RESOL2 IS 0.807

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU- BEST LINK	SIM- BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	1200
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	
	1	0.8843	10	0.91	0.867	0.016	0.8502	15	0.82	39	
	6	0.8709	10	0.90	0.851	0.016	0.8352	15	0.80	42	
	11	0.8619	1	0.88	0.844	0.007	0.8367	6	0.75	49	
	2	0.8286	15	0.86	0.745	0.098	0.6472	1	0.67	65	1200
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	
	9	0.8667	19	0.92	0.853	-0.027	0.8803	5	0.76	44	
	20	0.8496	19	0.92	0.807	0.047	0.7604	5	0.68	62	
	6	0.8058	20	0.82	0.674	0.133	0.5414	5	0.52	105	1204

POINTS NOT PLACED IN CLUSTERS
2 4

CLUSTERS AT RESOLUTION LEVEL 6 RESOL1 IS 0.637
RESOL2 IS 0.778

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	1200
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	
	1	0.8843	10	0.91	0.867	0.016	0.8502	15	0.82	39	
	6	0.8709	10	0.90	0.851	0.016	0.8352	15	0.80	42	
	11	0.8619	1	0.88	0.844	0.007	0.8367	6	0.75	49	
	2	0.8286	15	0.86	0.745	0.098	0.6472	1	0.67	65	
	3	0.7781	2	0.84	0.626	0.119	0.5075	1	0.52	109	1004
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	
	9	0.8667	19	0.92	0.853	-0.027	0.8803	5	0.76	44	
	20	0.8496	19	0.92	0.807	0.047	0.7604	5	0.68	62	
	6	0.8058	20	0.82	0.674	0.133	0.5414	5	0.52	105	1204

POINTS NOT PLACED IN CLUSTERS
4

CLUSTERS AT RESOLUTION LEVEL 7

RESOL1 IS 0.580
RESOL2 IS 0.750

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	
	15	0.7723	2	0.86	0.635	0.091	0.5440	18	0.48	123	1000
2	10										
	14	0.9403									
	15	0.9021	14	0.91	0.883	0.057	0.8257	10	0.86	27	
	1	0.8843	10	0.91	0.867	0.016	0.8502	15	0.82	39	
	6	0.8709	10	0.90	0.851	0.016	0.8352	15	0.80	42	
	11	0.8619	1	0.88	0.844	0.007	0.8367	6	0.75	49	
	2	0.8286	15	0.86	0.745	0.098	0.6472	1	0.67	65	4
3	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	
	9	0.8667	19	0.92	0.853	-0.027	0.8803	5	0.76	44	
	20	0.8496	19	0.92	0.807	0.047	0.7604	5	0.68	62	
	6	0.8058	20	0.82	0.674	0.133	0.5414	5	0.52	105	1204

POINTS NOT PLACED IN CLUSTERS
4

CLUSTERS AT RESOLUTION LEVEL 8 RESOL1 IS 0.523
RESOL2 IS 0.722

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	
	15	0.7723	2	0.86	0.635	0.091	0.5440	18	0.48	123	
	14	0.7308	15	0.91	0.606	0.029	0.5768	18	0.39	153	
	10	0.7009	14	0.94	0.596	0.010	0.5865	18	0.34	167	
	1	0.6791	10	0.91	0.592	0.004	0.5881	18	0.30	172	
	6	0.6698	10	0.90	0.628	-0.035	0.6631	18	0.31	170	
	11	0.6713	1	0.88	0.679	-0.052	0.7309	18	0.40	151	
	20	0.6592	6	0.82	0.592	0.087	0.5055	18	0.32	169	1000
2	16										
	17	0.9403									
	8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	
	5	0.9043	16	0.92	0.888	0.022	0.8660	8	0.84	37	
	19	0.8732	8	0.91	0.827	0.061	0.7652	5	0.75	50	
	9	0.8667	19	0.92	0.853	-0.027	0.8803	5	0.76	44	
	20	0.8496	19	0.92	0.807	0.047	0.7604	5	0.68	62	
	6	0.8058	20	0.82	0.674	0.133	0.5414	5	0.52	105	4

POINTS NOT PLACED IN CLUSTERS
4

CLUSTERS AT RESOLUTION LEVEL 9

RESOL1 IS 0.467
RESOL2 IS 0.693

CLUSTR AND MEMBER	AVGSIM OF CLUSTR	OTU- BEST LINK	SIM- BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1 7										
13	0.9437									
12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	
15	0.7723	2	0.86	0.635	0.091	0.5440	18	0.48	123	
14	0.7308	15	0.91	0.606	0.029	0.5768	18	0.39	153	
10	0.7009	14	0.94	0.596	0.010	0.5865	18	0.34	167	
1	0.6791	10	0.91	0.592	0.004	0.5881	18	0.30	172	
6	0.6698	10	0.90	0.628	-0.035	0.6631	18	0.31	170	
11	0.6713	1	0.88	0.679	-0.052	0.7309	18	0.40	151	
20	0.6592	6	0.82	0.592	0.087	0.5055	18	0.32	169	
19	0.6455	20	0.92	0.563	0.029	0.5346	18	0.29	175	
9	0.6348	19	0.92	0.565	-0.001	0.5664	18	0.23	181	
8	0.6231	19	0.91	0.542	0.023	0.5180	18	0.21	184	
17	0.6085	8	0.92	0.499	0.043	0.4560	18	0.16	187	1000
2 16										
17	0.9403									
8	0.9203	17	0.92	0.910	0.030	0.8805	16	0.90	15	4

POINTS NOT PLACED IN CLUSTERS
4 5

CLUSTERS AT RESOLUTION LEVEL 10

RESOL1 IS 0.410
RESOL2 IS 0.665

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	
	15	0.7723	2	0.86	0.635	0.091	0.5440	18	0.48	123	
	14	0.7308	15	0.91	0.606	0.029	0.5768	18	0.39	153	
	10	0.7009	14	0.94	0.596	0.010	0.5865	18	0.34	167	
	1	0.6791	10	0.91	0.592	0.004	0.5881	18	0.30	172	
	6	0.6698	10	0.90	0.628	-0.035	0.6631	18	0.31	170	
	11	0.6713	1	0.88	0.679	-0.052	0.7309	18	0.40	151	
	20	0.6592	6	0.82	0.592	0.087	0.5055	18	0.32	169	
	19	0.6455	20	0.92	0.563	0.029	0.5346	18	0.29	175	
	9	0.6348	19	0.92	0.565	-0.001	0.5664	18	0.23	181	
	8	0.6231	19	0.91	0.542	0.023	0.5180	18	0.21	184	
	17	0.6085	8	0.92	0.499	0.043	0.4560	18	0.16	187	
	16	0.5973	17	0.94	0.508	-0.009	0.5170	18	0.11	189	
	5	0.5841	16	0.92	0.471	0.037	0.4348	18	0.07	190	
	4	0.5775	3	0.64	0.518	-0.047	0.5649	1	0.32	168	230

POINTS NOT PLACED IN CLUSTERS
4

CLUSTERS AT RESOLUTION LEVEL 11 RESOL1 IS 0.353
RESOL2 IS 0.637

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	
	15	0.7723	2	0.86	0.635	0.091	0.5440	18	0.48	123	
	14	0.7308	15	0.91	0.606	0.029	0.5768	18	0.39	153	
	10	0.7009	14	0.94	0.596	0.010	0.5865	18	0.34	167	
	1	0.6791	10	0.91	0.592	0.004	0.5881	18	0.30	172	
	6	0.6698	10	0.90	0.628	-0.035	0.6631	18	0.31	170	
	11	0.6713	1	0.88	0.679	-0.052	0.7309	18	0.40	151	
	20	0.6592	6	0.82	0.592	0.087	0.5055	18	0.32	169	
	19	0.6455	20	0.92	0.563	0.029	0.5346	18	0.29	175	
	9	0.6348	19	0.92	0.565	-0.001	0.5664	18	0.23	181	
	8	0.6231	19	0.91	0.542	0.023	0.5180	18	0.21	184	
	17	0.6085	8	0.92	0.499	0.043	0.4560	18	0.16	187	
	16	0.5973	17	0.94	0.508	-0.009	0.5170	18	0.11	189	
	5	0.5841	16	0.92	0.471	0.037	0.4348	18	0.07	190	
	4	0.5775	3	0.64	0.518	-0.047	0.5649	1	0.32	168	230

POINTS NOT PLACED IN CLUSTERS
4

CLUSTERS AT RESOLUTION LEVEL 12

RESOL1 IS 0.297
RESOL2 IS 0.608

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	7										
	13	0.9437									
	12	0.9241	7	0.93	0.914	0.029	0.8849	13	0.90	19	
	3	0.9014	7	0.91	0.879	0.036	0.8429	12	0.84	31	
	18	0.8775	12	0.90	0.842	0.037	0.8050	3	0.76	47	
	2	0.8272	3	0.84	0.726	0.115	0.6111	18	0.61	78	
	15	0.7723	2	0.86	0.635	0.091	0.5440	18	0.48	123	
	14	0.7308	15	0.91	0.606	0.029	0.5768	18	0.39	153	
	10	0.7009	14	0.94	0.596	0.010	0.5865	18	0.34	167	
	1	0.6791	10	0.91	0.592	0.004	0.5881	18	0.30	172	
	6	0.6698	10	0.90	0.628	-0.035	0.6631	18	0.31	170	
	11	0.6713	1	0.88	0.679	-0.052	0.7309	18	0.40	151	
	20	0.6592	6	0.82	0.592	0.087	0.5055	18	0.32	169	
	19	0.6455	20	0.92	0.563	0.029	0.5346	18	0.29	175	
	9	0.6348	19	0.92	0.565	-0.001	0.5664	18	0.23	181	
	8	0.6231	19	0.91	0.542	0.023	0.5180	18	0.21	184	
	17	0.6085	8	0.92	0.499	0.043	0.4560	18	0.16	187	
	16	0.5973	17	0.94	0.508	-0.009	0.5170	18	0.11	189	
	5	0.5841	16	0.92	0.471	0.037	0.4348	18	0.07	190	
	4	0.5775	3	0.64	0.518	-0.047	0.5649	1	0.32	168	230

POINTS NOT PLACED IN CLUSTERS
4

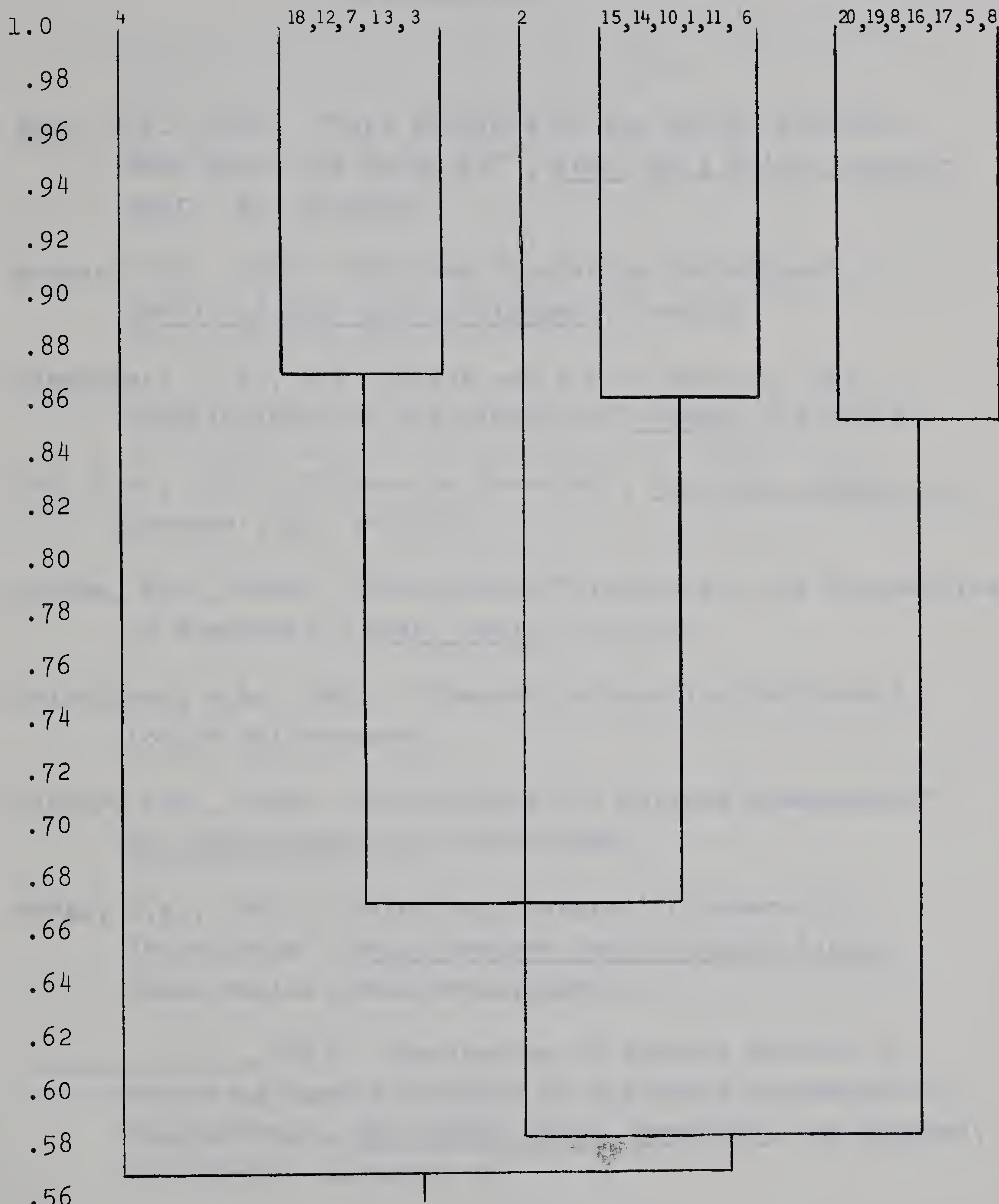
4.4 Interpretation of the Results

Since the points are not in very 'tight' clusters, the first level of interest is level 4 at which clusters 1, 2 and 3 become well defined. That is, there is a marked decrease in the value DIFF for the point responsible for termination of the cluster. Since it is obvious from the results that OTU 2 is a candidate for admission to both clusters 1 and 2 (see Figure 11) and is about the same distance away from the closest point in either one, it would seem reasonable not to include it in either one until clusters 1 and 2 are completely merged. Hence, the contents of cluster 1 and 2 can be ignored at levels 6 and 7. Cluster 3 is very stable since the addition of another OTU at levels 4, 5, 6 and 7 would violate three of the four criteria. When a cluster, or clusters, which have been formed at a higher level of resolution are taken apart at a lower one, this is an indication that we are in a transitional state, and the part-clusters which are formed should be ignored. Hence, level 9 should be ignored. The partition indicated at level 10 is very stable and remained this way for the remainder of the procedure levels. Note that OTU 4 is prevented from joining the cluster because of criteria 2 and 3. The

operation of these two criteria is better illustrated by the examples in the appendix. It should be kept in mind, however, that the most important criterion is the average linkage one since this gives us an indication of the relative 'goodness' of the custer. The other criteria are essential only to prevent additions of points when non-spherical, noisy, and/or unorthodox configurations arise.

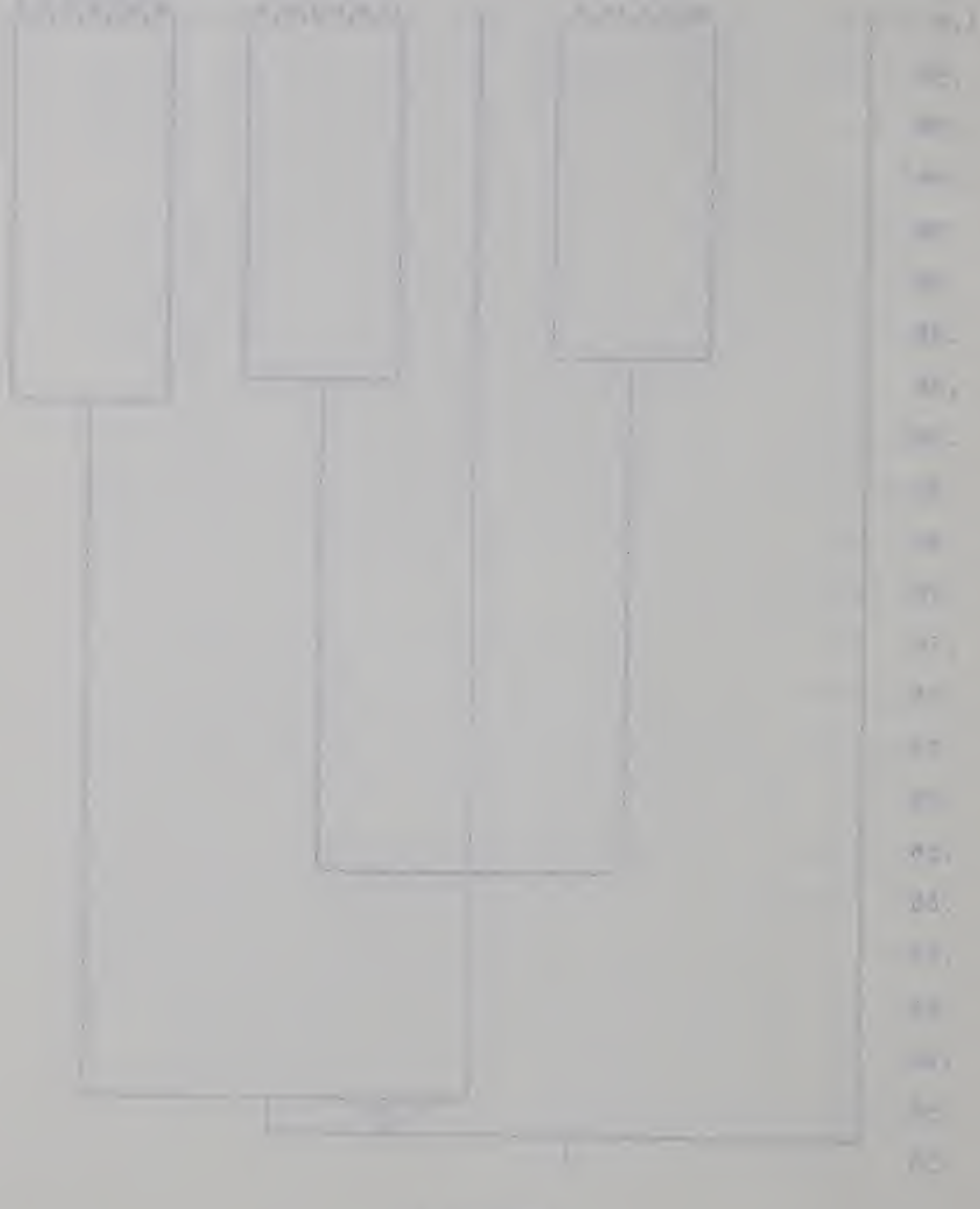
For this simple example, the one dimensional arrangement in the dendogram is not too bad a summary of the interpoint relations. However, if the points are not roughly unidimensional, the dendogram may be a very poor representation.

One possibility is to attempt to find a three dimensional configuration for the clusters using multi-dimensional scaling techniques such as those proposed by Shephard (1962) and Kruskal (1964). These are iterative procedures for fitting OTU's into spaces of successively lower dimension. If the three dimensional configuration is not 'too' distorted, a physical model of the configuration with the distortion of each link represented in some way might be a good description of the relationships existing in the population.



1.0
.98
.96
.94
.92
.90
.88
.86
.84
.82
.80
.78
.76
.74
.72
.70
.68
.66
.64
.62
.60
.58
.56
4
18, 12, 7, 13, 3
2
15, 14, 10, 1, 11, 6
20, 19, 8, 16, 17, 5, 8

Figure 13



BIBLIOGRAPHY

Ball, G.H., 1965. "Data Analysis in the Social Sciences: What About the Details?" , Proc. Fall Joint Computer Conf., pp. 533-559.

Bonner, R.E., 1964. "On Some Clustering Techniques", IBM J. of Res. and Development, January.

Carmichael, J.W., R.S. Julius and P.M.D. Martin, 1965. "Similarities in One Dimension", Nature 208:544-547.

Cox, D.R., 1957. "A Note on Grouping", Am. Stat. Assoc. J., December, pp. 543-545.

Dupraw, E.J., 1965. "Non-Linnean Taxonomy and the Systematics of Honeybees", Syst. Zool., 14:1-24.

Fairthorne, R.A., 1961. "Towards Information Retrieval", London Butterworths.

Fisher, W.D., 1958. "On Grouping for Maximum Homogeneity", Am. Stat. Assoc. J., 53:789-798.

Forgy, E.W., 1963. "Detecting 'Natural' Clusters of Individuals", Proc. Western Psychological Assoc., Santa Monica, California, April.

______________, 1964. "Evaluation of Several Methods of Detecting Sample Mixtures of Different N-Dimensional Populations", Am. Psych. Assoc. Meetings, Los Angeles, California, September 9.

________________, 1965. "Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications", AAAS-Biometric Society (WNAR) Meeting, Riverside, California, June 22.

Hill, L.R., L.G. Silvestri, P. Ihm, G. Farchi and P. Lanciani, 1965. "Automatic Classification of Staphylococci by Principle Component Analysis and a Gradient Method", J. of Bacteriology, pp. 1393-1401.

Johnson, S.C., "Hierarchical Clustering Schemes", Bell Telephones Laboratories Internal Publication.

Kruskal, J.B., 1964. "Multidimensional Scaling by Optimizing Goodness of Fit to a Non-metric Hypothesis", Psychometrika 29:1-27.

________________, 1964. "Multidimensional Scaling: A Numerical Method", Psychometrika, 29:115-129.

Macnaughton-Smith, P., W.T. Williams, M.B. Dale and L.G. Mockett, 1964. "Dissimilarity Analysis: A New Technique of Hierarchical Subdivision", Nature 202:1034-5.

Mattson, R.L. and J.E. Dammann, 1965. "A Technique for Determining and Coding Subclasses in Pattern Recognition Problems", IBM J. of Res. and Develop., July, pp. 294-302.

Mitchener, C.D. and R.R. Sokal, 1957. "A Quantitative Approach to a Problem in Classification", Evolution 11:130-162.

Rogers, D.J. and T. T. Tanimoto, 1960. "A Computer Program for Classifying Plants", Science 132:1115-1118.

Rohlf, F.J. and R. R. Sokal, 1962. "The Description of Taxonomic Relationships by Factor Analysis", Syst. Zool. 11:1-16.

Rohlf, F.J., 1965. "Multivariate Methods in Taxonomy", Proc. Fall Joint Comp. Conf.

Schuessler, K.F. and H. Driver, 1956. "A Factor Analysis of Sixteen Primitive Societies", Am. Sociol. Rev. 21:493-499.

Shephard, R.N., 1962. "The Analysis of Proximities: A Multidimensional Scaling with an Unknown Distance Function.I.", Psychometrika 27:125-140.

_______________, 1962. "The Analysis of Proximities: A Multidimensional Scaling with an Unknown Distance Function.II.", Psychometrika 27:219-246.

Shephard, R.N. and J.D. Carrol, 1965. "Parametric Representation of Non-Linear Data Structures", Int. Symposium on Multivariate Analysis, June 14-19.

Silvestri, L., M. Turri, L.R. Hill and E. Gilardi, 1962. "A Quantitative Approach to the Systematics of Actinomycetes Based on Overall Similarity", Microbial Classification, 12th Symposium of the Society for General Microbiology, pp. 333-360.

Simpson, G.G., 1945. "The Principles of Classification and a Classification of Mammals", Bulletin Am. Museum of Natural History, 85:1-350.

Sneath, P.H.A., 1957. "Some Thoughts on Bacterial Classification", J. Gen. Microbiol. 17:184-200.

________________, 1957. "The Application of Computers to Taxonomy", J. Gen. Microbiol. 17:201-226.

________________, 1961. "Recent Development in Theoretical and Quantitiative Taxonomy", Syst. Zool. 10:118-139.

Sokal, R.R., 1961. "Distance as a Measure of Taxonomic Similarity", Syst. Zool. 10:70-79.

Sokal, R.R. and P.H.A. Sneath, 1963. "Principles of Numerical Taxonomy", W.H. Freeman and Co., San Francisco & London.

Stroud, C.P., 1953. "An Application of Factor Analysis to the Systematics of Kalotermes", Syst. Zool. 2:76-92.

Ward, J.H., 1963. "Hierarchical Grouping to Optimize an Objective Function", Am. Stat. Assoc. J., March, pp. 236-244.

Williams, W.T. and M.B. Dale, 1964. "An Objective Method of Weighing in Similarity Analysis", Nature 201:426.

__________________ and G.N. Lance, 1965. "Logic of Computer-Based Intrinsic Classification", Nature 207:159-161.

APPENDIX

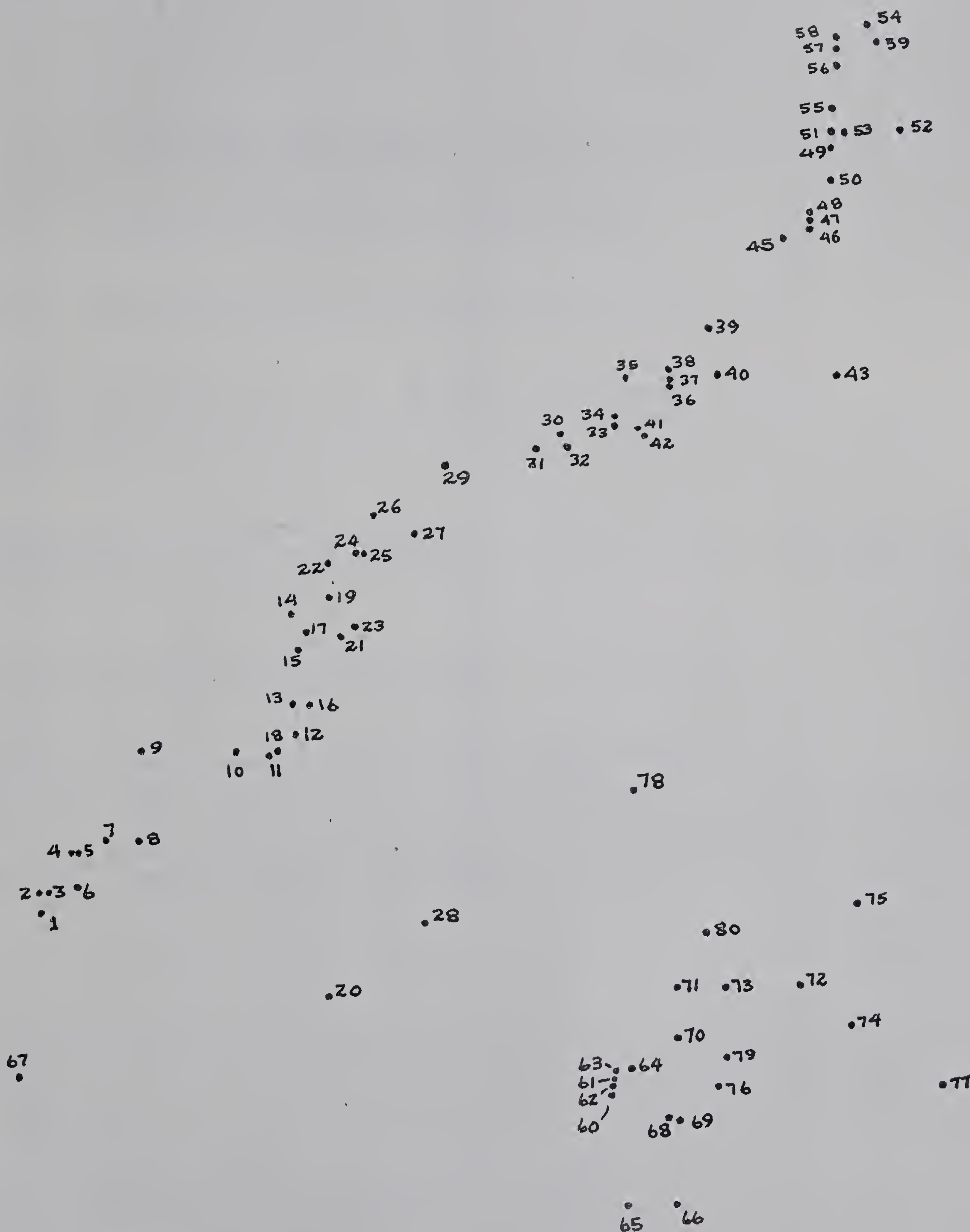
This appendix contains results of the application of the method to three special configurations of points. Only the results from the more informative levels for each configuration have been included. The similarities between the points were computed using the procedure suggested in Section 2.2.

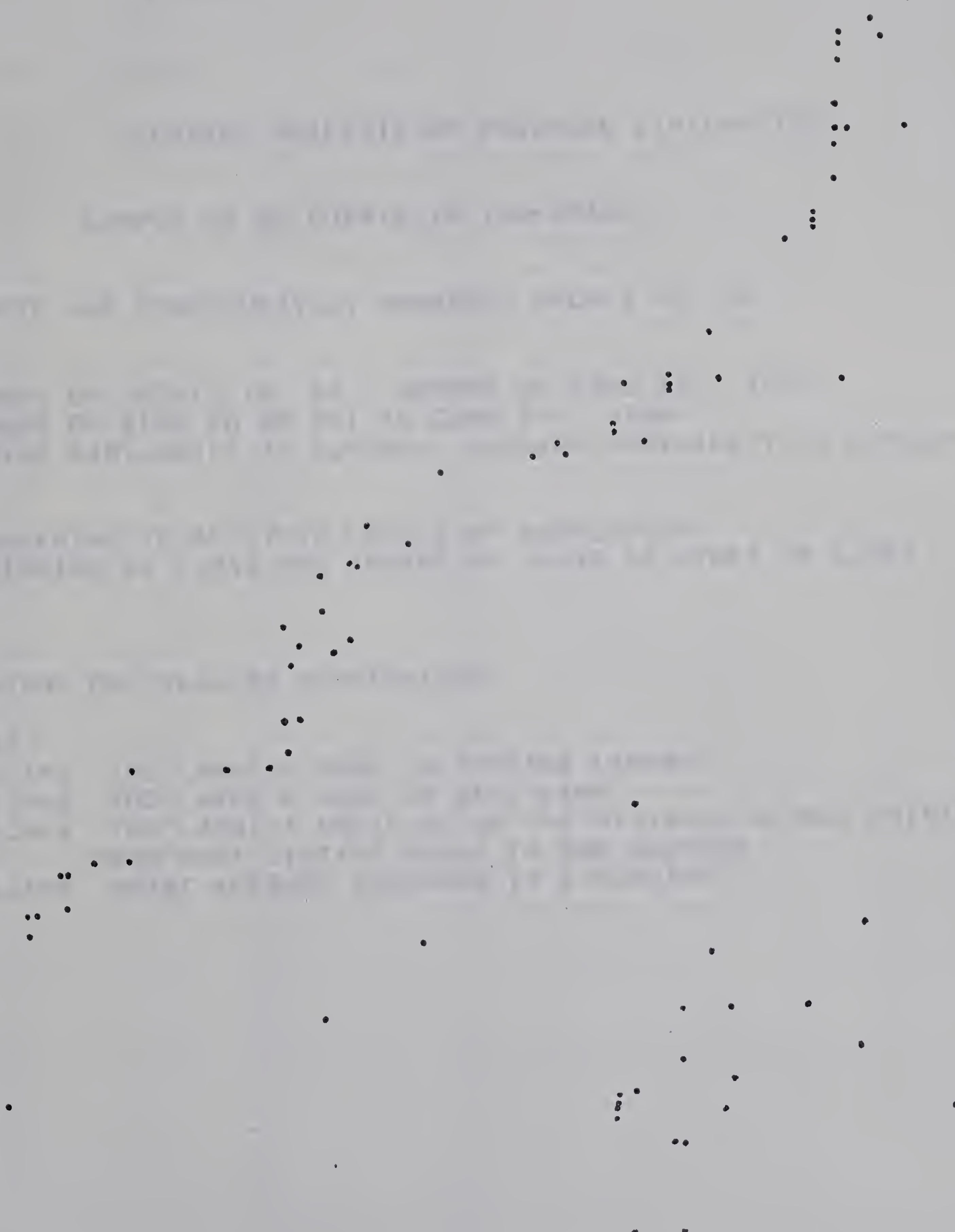
Configuration I consists of two obvious clusters, the elongate one consisting of four smaller ones. Upon application of the procedure at high resolution, the four smaller ones are detected along with the large globose one whereas at lower resolution only the two large clusters are displayed.

Configuration II consists of two parallel ellipsoidal clusters with some 'bridging' points between them. The procedure successfully detects these two clusters.

The third configuration of points demonstrates the capability of the procedure for detecting clusters which are well separated yet whose centroids are nearly concident. This emphasizes the importance of looking for a 'gap' or discontinuity rather than using absolute criteria as a basis for detecting clusters.

CONFIGURATION I





CLUSTER ANALYSIS OF RELATIVE SIMILARITIES

SAMPLE OF 80 POINTS IN TWO-SPACE

POINTS ARE CONSECUTIVELY NUMBERED FROM 1 TO 80

NUMBER OF POINTS IS 80 NUMBER OF SIMS IS 3160
NUMBER OF SIMS TO BE PUT IN CORE IS 3160
MEDIAN SIMILARITY IS 0.62304 AVERAGE SIMILARITY IS 0.64265

CLUSTERING TO BE AT 12 LEVELS OF RESOLUTION
BEGINNING AT 0.940 AND ENDING AT 0.372 BY STEPS OF 0.052

REASONS FOR CLUSTER TERMINATION

FLAG
COL.1=1 TOO LARGE A DROP IN AVERAGE LINKAGE
COL.2=2 TOO LARGE A DROP IN BEST LINK
COL.3=3 TOO LARGE A DECREASE IN THE DISTANCE OF NEW POINT
FROM MOST DISTANT POINT IN THE CLUSTER
COL.4=4 POINT ALREADY INCLUDED IN A CLUSTER

FREQUENCIES OF SIMILARITIES AT 0.01 LEVELS

LEVEL	NUMBER	TOTAL,	REMAINDER	HISTOGRAM
1.00	21	21	3139	*******
0.99	31	52	3108	**********
0.98	37	89	3071	************
0.97	48	137	3023	****************
0.96	56	193	2967	*******************
0.95	48	241	2919	****************
0.94	46	287	2873	***************
0.93	46	333	2827	***************
0.92	45	378	2782	***************
0.91	38	416	2744	*************
0.90	36	452	2708	************
0.89	32	484	2676	***********
0.88	39	523	2637	*************
0.87	39	562	2598	*************
0.86	33	595	2565	***********
0.85	37	632	2528	************
0.84	45	677	2483	***************
0.83	37	714	2446	************
0.82	46	760	2400	***************
0.81	48	808	2352	****************
0.80	39	847	2313	*************
0.79	48	895	2265	****************
0.78	49	944	2216	****************
0.77	50	994	2166	*****************
0.76	49	1043	2117	****************
0.75	44	1087	2073	***************
0.74	39	1126	2034	*************
0.73	39	1165	1995	*************
0.72	34	1199	1961	***********
0.71	32	1231	1929	***********
0.70	34	1265	1895	***********
0.69	40	1305	1855	*************
0.68	45	1350	1810	***************
0.67	41	1391	1769	**************
0.66	46	1437	1723	***************
0.65	46	1483	1677	***************
0.64	55	1538	1622	******************
0.63	60	1598	1562	********************
0.62	83	1681	1479	****************************
0.61	63	1744	1416	*********************
0.60	98	1842	1318	*********************************
0.59	51	1893	1267	*****************
0.58	92	1985	1175	*******************************
0.57	74	2059	1101	*************************
0.56	68	2127	1033	***********************
0.55	65	2192	968	**********************
0.54	79	2271	889	**************************
0.53	43	2314	846	**************
0.52	45	2359	801	***************
0.51	44	2403	757	***************

0.50	56	2459	701	******************
0.49	51	2510	650	*****************
0.48	39	2549	611	*************
0.47	32	2581	579	**********
0.46	37	2618	542	************
0.45	25	2643	517	********
0.44	32	2675	485	**********
0.43	28	2703	457	*********
0.42	36	2739	421	************
0.41	31	2770	390	**********
0.40	27	2797	363	*********
0.39	27	2824	336	*********
0.38	23	2847	313	*******
0.37	31	2878	282	**********
0.36	24	2902	258	********
0.35	28	2930	230	*********
0.34	26	2956	204	********
0.33	19	2975	185	******
0.32	24	2999	161	********
0.31	18	3017	143	******
0.30	21	3038	122	*******
0.29	17	3055	105	*****
0.28	21	3076	84	*******
0.27	20	3096	64	******
0.26	11	3107	53	***
0.25	4	3111	49	*
0.24	9	3120	40	***
0.23	10	3130	30	***
0.22	9	3139	21	***
0.21	6	3145	15	**
0.20	4	3149	11	*
0.19	1	3150	10	*
0.18	1	3151	9	*
0.17	2	3153	7	*
0.16	1	3154	6	*
0.15	0	3154	6	
0.14	1	3155	5	*
0.13	1	3156	4	*
0.12	2	3158	2	*
0.11	0	3158	2	
0.10	2	3160	-0	*
0.09	0	3160	-0	
0.08	0	3160	-0	
0.07	0	3160	-0	
0.06	0	3160	-0	
0.05	0	3160	-0	
0.04	0	3160	-0	
0.03	0	3160	-0	
0.02	0	3160	-0	
0.01	0	3160	-0	

CLUSTERS AT RESOLUTION LEVEL 3

RESOL1 IS 0.837
RESOL2 IS 0.888

CLUSTR AND	MEMBER	AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	11										
	18	0.9972									
	12	0.9863	11	0.98	0.981	0.016	0.9644	18	0.98	53	
	13	0.9786	12	0.98	0.971	0.010	0.9612	18	0.97	113	
	16	0.9768	13	0.99	0.974	-0.003	0.9770	18	0.96	134	
	10	0.9709	18	0.97	0.959	0.015	0.9442	16	0.94	228	
	15	0.9626	13	0.96	0.942	0.017	0.9246	10	0.92	332	
	17	0.9565	15	0.99	0.938	0.004	0.9345	10	0.91	386	
	14	0.9518	17	0.99	0.936	0.003	0.9330	10	0.90	412	
	21	0.9494	17	0.98	0.940	-0.004	0.9441	10	0.90	429	
	23	0.9468	21	0.99	0.935	0.005	0.9297	10	0.88	474	
	19	0.9446	21	0.97	0.934	0.001	0.9332	10	0.88	486	
	22	0.9407	19	0.98	0.919	0.015	0.9038	10	0.86	562	
	24	0.9369	22	0.98	0.914	0.005	0.9092	10	0.85	612	
	25	0.9343	24	0.99	0.917	-0.003	0.9205	10	0.84	623	
	26	0.9298	25	0.97	0.898	0.019	0.8795	10	0.82	717	
	27	0.9258	26	0.97	0.896	0.002	0.8941	10	0.81	755	
	29	0.9177	27	0.95	0.853	0.043	0.8094	10	0.76	991	1000
2	4										
	5	0.9972									
	7	0.9856	5	0.98	0.980	0.017	0.9624	4	0.98	56	
	6	0.9794	5	0.98	0.973	0.007	0.9665	7	0.96	116	
	3	0.9737	6	0.98	0.965	0.008	0.9570	7	0.95	201	
	2	0.9717	3	0.99	0.968	-0.003	0.9703	7	0.94	218	
	1	0.9697	2	0.99	0.965	0.003	0.9621	7	0.94	259	
	8	0.9631	7	0.97	0.943	0.022	0.9212	1	0.92	347	
	9	0.9506	8	0.94	0.907	0.036	0.8713	1	0.87	511	
	10	0.9354	9	0.94	0.875	0.033	0.8420	1	0.83	669	4
3	37										
	38	0.9972									
	36	0.9962	37	1.00	0.996	0.001	0.9944	38	0.99	12	
	35	0.9823	38	0.97	0.968	0.027	0.9410	36	0.97	101	
	34	0.9749	35	0.98	0.964	0.005	0.9593	38	0.96	145	
	33	0.9717	34	0.99	0.965	-0.001	0.9666	38	0.95	173	
	41	0.9714	33	0.99	0.971	-0.005	0.9762	38	0.96	133	
	42	0.9710	41	0.99	0.970	0.001	0.9690	35	0.96	149	
	40	0.9654	38	0.97	0.946	0.024	0.9212	33	0.92	317	
	39	0.9594	40	0.97	0.936	0.010	0.9257	33	0.91	377	
	32	0.9539	33	0.96	0.929	0.007	0.9226	39	0.88	494	
	30	0.9501	32	0.99	0.931	-0.002	0.9329	39	0.87	502	

	31	0.9457	30	0.98	0.921	0.010	0.9114	39	0.86	566	
	29	0.9351	31	0.94	0.872	0.050	0.8220	39	0.81	780	1000

4	47										
	48	0.9972									
	46	0.9925	47	0.99	0.990	0.007	0.9831	48	0.99	24	
	44	0.9899	46	0.99	0.987	0.003	0.9846	48	0.98	42	
	45	0.9856	44	0.99	0.979	0.008	0.9710	48	0.97	79	
	50	0.9787	48	0.98	0.965	0.014	0.9506	45	0.95	196	
	49	0.9713	50	0.98	0.953	0.012	0.9402	45	0.93	273	
	51	0.9656	49	0.99	0.948	0.004	0.9444	45	0.92	323	
	53	0.9624	51	0.99	0.952	-0.003	0.9547	45	0.92	336	
	55	0.9586	51	0.98	0.943	0.008	0.9346	45	0.91	389	
	56	0.9516	55	0.97	0.920	0.023	0.8977	45	0.88	488	
	57	0.9458	56	0.99	0.916	0.004	0.9126	45	0.87	532	
	58	0.9415	57	0.99	0.918	-0.001	0.9192	45	0.86	553	
	54	0.9368	58	0.98	0.909	0.009	0.9002	45	0.85	609	
	59	0.9346	54	0.99	0.920	-0.011	0.9309	45	0.85	582	
	52	0.9341	53	0.96	0.931	-0.011	0.9415	45	0.89	439	
	39	0.9237	45	0.92	0.846	0.085	0.7610	54	0.77	938	1204

5	68										
	69	0.9972									
	76	0.9744	69	0.96	0.963	0.034	0.9288	68	0.96	130	
	79	0.9665	76	0.98	0.959	0.004	0.9544	68	0.95	211	
	70	0.9611	79	0.96	0.953	0.006	0.9471	68	0.94	223	
	71	0.9513	70	0.97	0.932	0.021	0.9108	68	0.91	379	
	73	0.9472	71	0.97	0.937	-0.005	0.9422	68	0.90	398	
	64	0.9461	70	0.96	0.943	-0.006	0.9487	73	0.92	343	
	63	0.9455	64	0.99	0.943	-0.000	0.9432	73	0.91	376	
	61	0.9460	63	0.99	0.948	-0.005	0.9532	73	0.91	385	
	62	0.9471	61	1.00	0.952	-0.004	0.9565	73	0.91	396	
	60	0.9478	62	0.99	0.951	0.001	0.9499	73	0.90	424	
	80	0.9410	73	0.96	0.903	0.048	0.8558	68	0.87	519	
	72	0.9342	73	0.95	0.893	0.010	0.8831	60	0.86	575	
	74	0.9263	72	0.96	0.875	0.018	0.8574	60	0.83	653	
	66	0.9213	68	0.94	0.886	-0.011	0.8966	80	0.82	728	
	65	0.9172	66	0.97	0.887	-0.001	0.8878	74	0.81	771	
	75	0.9084	72	0.94	0.838	0.049	0.7888	65	0.75	1051	1000

POINTS NOT PLACED IN CLUSTERS

20 28 29 43 67 75 77 78

CLUSTERS AT RESOLUTION LEVEL 10 RESOL1 IS 0.475
RESOL2 IS 0.707

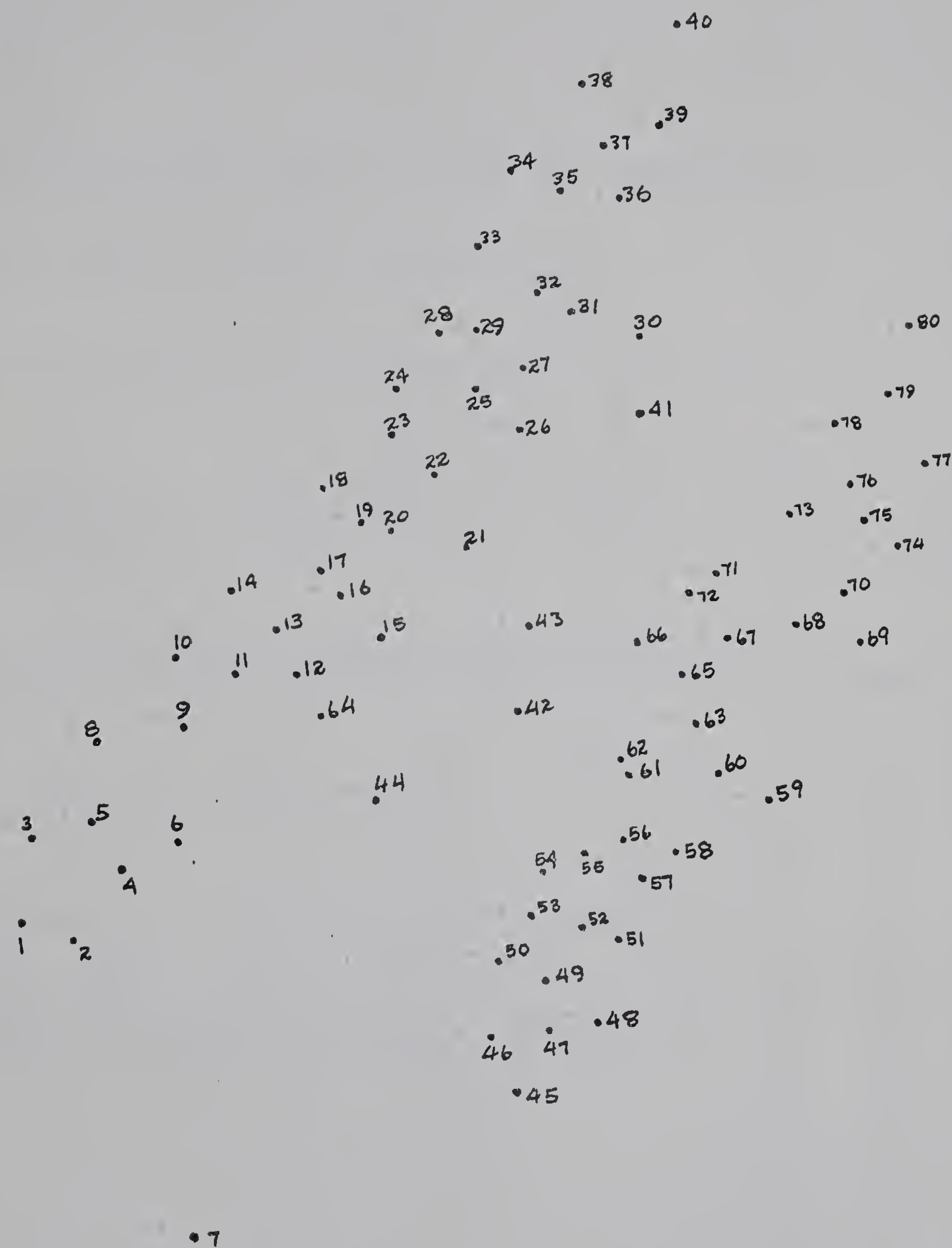
CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	11										
	18	0.9972									
	12	0.9863	11	0.98	0.981	0.016	0.9644	18	0.98	53	
	13	0.9786	12	0.98	0.971	0.010	0.9612	18	0.97	113	
	16	0.9768	13	0.99	0.974	-0.003	0.9770	18	0.96	134	
	10	0.9709	18	0.97	0.959	0.015	0.9442	16	0.94	228	
	15	0.9626	13	0.96	0.942	0.017	0.9246	10	0.92	332	
	17	0.9565	15	0.99	0.938	0.004	0.9345	10	0.91	386	
	14	0.9518	17	0.99	0.936	0.003	0.9330	10	0.90	412	
	21	0.9494	17	0.98	0.940	-0.004	0.9441	10	0.90	429	
	23	0.9468	21	0.99	0.935	0.005	0.9297	10	0.88	474	
	19	0.9446	21	0.97	0.934	0.001	0.9332	10	0.88	486	
	22	0.9407	19	0.98	0.919	0.015	0.9038	10	0.86	562	
	24	0.9369	22	0.98	0.914	0.005	0.9092	10	0.85	612	
	25	0.9343	24	0.99	0.917	-0.003	0.9205	10	0.84	623	
	26	0.9298	25	0.97	0.898	0.019	0.8795	10	0.82	717	
	27	0.9258	26	0.97	0.896	0.002	0.8941	10	0.81	755	
	29	0.9177	27	0.95	0.853	0.043	0.8094	10	0.76	991	
	31	0.9069	29	0.94	0.815	0.038	0.7775	10	0.72	1178	
	30	0.8971	31	0.98	0.809	0.006	0.8024	10	0.70	1229	
	32	0.8894	30	0.99	0.816	-0.007	0.8227	10	0.70	1230	
	33	0.8804	32	0.96	0.790	0.026	0.7645	10	0.67	1367	
	34	0.8729	33	0.99	0.795	-0.005	0.7997	10	0.66	1390	
	41	0.8666	33	0.99	0.797	-0.002	0.7996	10	0.66	1407	
	42	0.8615	41	0.99	0.803	-0.005	0.8081	10	0.66	1412	
	35	0.8567	34	0.98	0.799	0.003	0.7963	10	0.64	1475	
	38	0.8515	35	0.97	0.786	0.013	0.7726	10	0.62	1603	
	37	0.8475	38	1.00	0.795	-0.009	0.8046	10	0.62	1588	
	36	0.8445	37	1.00	0.804	-0.008	0.8123	10	0.62	1579	
	40	0.8401	38	0.97	0.779	0.025	0.7541	10	0.59	1824	
	39	0.8358	40	0.97	0.773	0.006	0.7662	10	0.58	1929	
	9	0.8300	10	0.94	0.744	0.029	0.7148	39	0.53	2278	
	8	0.8230	9	0.94	0.715	0.029	0.6861	39	0.49	2457	
	7	0.8161	8	0.97	0.705	0.010	0.6953	39	0.47	2541	
	5	0.8092	7	0.98	0.696	0.010	0.6860	39	0.45	2602	
	4	0.8032	5	1.00	0.702	-0.007	0.7086	39	0.45	2615	
	6	0.7975	5	0.98	0.697	0.005	0.6923	39	0.44	2649	
	3	0.7917	6	0.98	0.687	0.010	0.6778	39	0.42	2703	
	2	0.7866	3	0.99	0.691	-0.004	0.6954	39	0.42	2715	
	1	0.7818	2	0.99	0.692	-0.000	0.6921	39	0.41	2746	
	45	0.7744	39	0.92	0.630	0.062	0.5677	1	0.33	2952	
	44	0.7674	45	0.99	0.626	0.003	0.6232	1	0.32	2977	
	46	0.7610	44	0.99	0.630	-0.003	0.6326	1	0.31	2994	
	47	0.7551	46	0.99	0.632	-0.003	0.6351	1	0.31	3002	

	48	0.7499	47	1.00	0.639	-0.006	0.6451	1	0.31	3008	
	50	0.7444	48	0.98	0.623	0.016	0.6073	1	0.28	3054	
	49	0.7390	50	0.98	0.617	0.006	0.6103	1	0.27	3083	
	51	0.7339	49	0.99	0.616	0.000	0.6157	1	0.26	3095	
	53	0.7292	51	0.99	0.619	-0.003	0.6215	1	0.25	3105	
	55	0.7247	51	0.98	0.618	0.000	0.6181	1	0.25	3108	
	56	0.7199	55	0.97	0.600	0.018	0.5822	1	0.22	3127	
	57	0.7152	56	0.99	0.599	0.001	0.5974	1	0.22	3133	
	58	0.7109	57	0.99	0.602	-0.003	0.6049	1	0.21	3138	
	54	0.7065	58	0.98	0.590	0.012	0.5785	1	0.19	3149	
	59	0.7027	54	0.99	0.602	-0.012	0.6146	1	0.19	3148	
	52	0.7003	53	0.96	0.635	-0.033	0.6678	1	0.23	3124	
	43	0.7003	40	0.92	0.702	-0.067	0.7692	1	0.36	2885	30
2	68										
	69	0.9972									
	76	0.9744	69	0.96	0.963	0.034	0.9288	68	0.96	130	
	79	0.9665	76	0.98	0.959	0.004	0.9544	68	0.95	211	
	70	0.9611	79	0.96	0.953	0.006	0.9471	68	0.94	223	
	71	0.9513	70	0.97	0.932	0.021	0.9108	68	0.91	379	
	73	0.9472	71	0.97	0.937	-0.005	0.9422	68	0.90	398	
	64	0.9461	70	0.96	0.943	-0.006	0.9487	73	0.92	343	
	63	0.9455	64	0.99	0.943	-0.000	0.9432	73	0.91	376	
	61	0.9460	63	0.99	0.948	-0.005	0.9532	73	0.91	385	
	62	0.9471	61	1.00	0.952	-0.004	0.9565	73	0.91	396	
	60	0.9478	62	0.99	0.951	0.001	0.9499	73	0.90	424	
	80	0.9410	73	0.96	0.903	0.048	0.8558	68	0.87	519	
	72	0.9342	73	0.95	0.893	0.010	0.8831	60	0.86	575	
	74	0.9263	72	0.96	0.875	0.018	0.8574	60	0.83	653	
	66	0.9213	68	0.94	0.886	-0.011	0.8966	80	0.82	728	
	65	0.9172	66	0.97	0.887	-0.001	0.8878	74	0.81	771	
	75	0.9084	72	0.94	0.838	0.049	0.7888	65	0.75	1051	
	77	0.8994	74	0.93	0.823	0.015	0.8084	65	0.78	912	
	78	0.8897	80	0.89	0.802	0.021	0.7803	77	0.71	1183	
	28	0.8797	63	0.84	0.785	0.017	0.7681	77	0.64	1497	
	20	0.8680	28	0.93	0.751	0.034	0.7170	77	0.59	1843	
	11	0.8504	28	0.85	0.665	0.086	0.5798	77	0.50	2386	4

POINTS NOT PLACED IN CLUSTERS

43 67

CONFIGURATION II



CLUSTER ANALYSIS OF RELATIVE SIMILARITIES

SAMPLE OF 80 POINTS IN TWO-SPACE

POINTS ARE CONSECUTIVELY NUMBERED FROM 1 TO 80

NUMBER OF POINTS IS 80 NUMBER OF SIMS IS 3160
NUMBER OF SIMS TO BE PUT IN CORE IS 3160
MEDIAN SIMILARITY IS 0.71513 AVERAGE SIMILARITY IS 0.71090

CLUSTERING TO BE AT 12 LEVELS OF RESOLUTION
BEGINNING AT 0.920 AND ENDING AT 0.507 BY STEPS OF 0.037

REASONS FOR CLUSTER TERMINATION

FLAG
COL.1=1 TOO LARGE A DROP IN AVERAGE LINKAGE
COL.2=2 TOO LARGE A DROP IN BEST LINK
COL.3=3 TOO LARGE A DECREASE IN THE DISTANCE OF NEW POINT
FROM MOST DISTANT POINT IN THE CLUSTER
COL.4=4 POINT ALREADY INCLUDED IN A CLUSTER

FREQUENCIES OF SIMILARITIES AT 0.01 LEVELS

LEVEL	NUMBER	TOTAL,	REMAINDER	HISTOGRAM
1.00	0	0	3160	
0.99	1	1	3159	*
0.98	11	12	3148	***
0.97	27	39	3121	*********
0.96	48	87	3073	****************
0.95	44	131	3029	**************
0.94	58	189	2971	*******************
0.93	46	235	2925	***************
0.92	53	288	2872	*****************
0.91	45	333	2827	***************
0.90	56	389	2771	******************
0.89	56	445	2715	******************
0.88	41	486	2674	*************
0.87	59	545	2615	*******************
0.86	62	607	2553	********************
0.85	49	656	2504	****************
0.84	59	715	2445	*******************
0.83	57	772	2388	*******************
0.82	56	828	2332	******************
0.81	71	899	2261	***********************
0.80	63	962	2198	*********************
0.79	54	1016	2144	******************
0.78	78	1094	2066	**************************
0.77	91	1185	1975	******************************
0.76	83	1268	1892	***************************
0.75	68	1336	1824	**********************
0.74	100	1436	1724	*********************************
0.73	101	1537	1623	*********************************
0.72	83	1620	1540	***************************
0.71	91	1711	1449	******************************
0.70	119	1830	1330	***************************************
0.69	71	1901	1259	***********************
0.68	92	1993	1167	******************************
0.67	96	2089	1071	********************************
0.66	92	2181	979	******************************
0.65	77	2258	902	*************************
0.64	67	2325	835	**********************
0.63	66	2391	769	**********************
0.62	65	2456	704	*********************
0.61	62	2518	642	********************
0.60	44	2562	598	**************
0.59	45	2607	553	***************
0.58	50	2657	503	****************
0.57	45	2702	458	***************
0.56	33	2735	425	***********
0.55	33	2768	392	***********
0.54	44	2812	348	**************
0.53	31	2843	317	**********
0.52	28	2871	289	*********
0.51	25	2896	264	********

0.50	25	2921	239	********
0.49	23	2944	216	*******
0.48	23	2967	193	*******
0.47	14	2981	179	****
0.46	22	3003	157	*******
0.45	18	3021	139	******
0.44	17	3038	122	*****
0.43	11	3049	111	***
0.42	8	3057	103	**
0.41	19	3076	84	******
0.40	15	3091	69	*****
0.39	7	3098	62	**
0.38	9	3107	53	***
0.37	10	3117	43	***
0.36	6	3123	37	**
0.35	4	3127	33	*
0.34	9	3136	24	***
0.33	5	3141	19	*
0.32	4	3145	15	*
0.31	1	3146	14	*
0.30	2	3148	12	*
0.29	1	3149	11	*
0.28	2	3151	9	*
0.27	3	3154	6	*
0.26	1	3155	5	*
0.25	0	3155	5	
0.24	0	3155	5	
0.23	2	3157	3	*
0.22	0	3157	3	
0.21	0	3157	3	
0.20	2	3159	1	*
0.19	0	3159	1	
0.18	0	3159	1	
0.17	0	3159	1	
0.16	0	3159	1	
0.15	0	3159	1	
0.14	1	3160	-0	*
0.13	0	3160	-0	
0.12	0	3160	-0	
0.11	0	3160	-0	
0.10	0	3160	-0	
0.09	0	3160	-0	
0.08	0	3160	-0	
0.07	0	3160	-0	
0.06	0	3160	-0	
0.05	0	3160	-0	
0.04	0	3160	-0	
0.03	0	3160	-0	
0.02	0	3160	-0	
0.01	0	3160	-0	

CLUSTERS AT RESOLUTION LEVEL 12

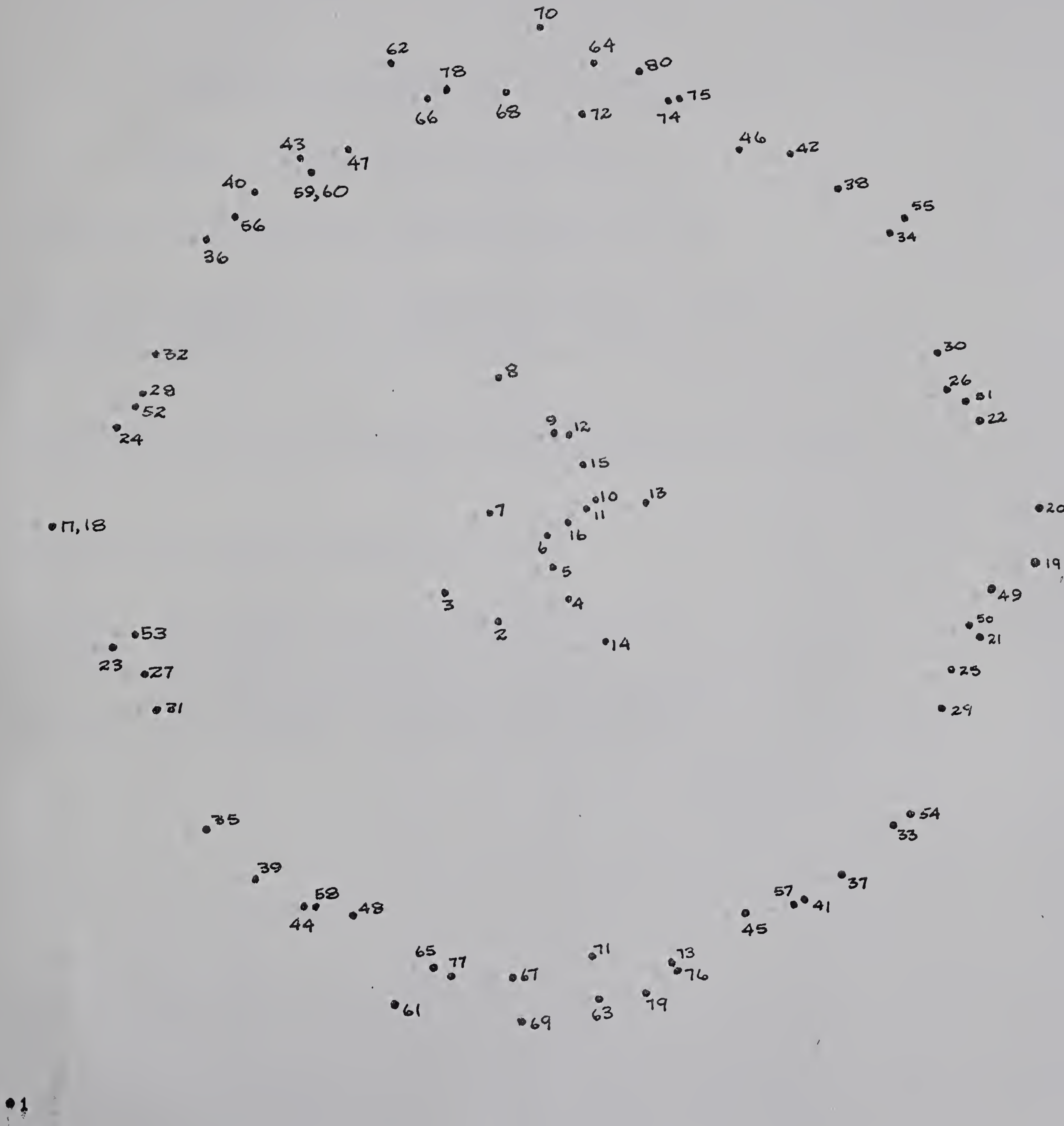
RESOL1 IS 0.507
RESOL2 IS 0.714

CLUSTR AND MEMBER		AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	61										
	62	0.9894									
	56	0.9640	61	0.96	0.951	0.038	0.9133	62	0.95	101	
	57	0.9532	56	0.97	0.942	0.009	0.9333	62	0.92	229	
	58	0.9531	57	0.97	0.953	-0.011	0.9637	62	0.93	178	
	55	0.9517	56	0.97	0.949	0.004	0.9445	62	0.93	166	
	54	0.9447	55	0.97	0.927	0.021	0.9058	62	0.91	302	
	53	0.9373	54	0.97	0.915	0.012	0.9029	62	0.88	453	
	52	0.9359	53	0.97	0.931	-0.016	0.9471	62	0.89	412	
	51	0.9346	52	0.97	0.929	0.002	0.9274	62	0.88	448	
	50	0.9281	53	0.96	0.899	0.030	0.8689	62	0.84	653	
	49	0.9253	50	0.97	0.911	-0.012	0.9237	62	0.84	643	
	47	0.9204	49	0.97	0.893	0.018	0.8747	62	0.82	797	
	48	0.9179	47	0.97	0.903	-0.010	0.9124	62	0.82	746	
	46	0.9134	47	0.96	0.885	0.018	0.8666	62	0.80	926	
	45	0.9080	46	0.96	0.870	0.015	0.8547	62	0.77	1094	
	63	0.9006	62	0.95	0.846	0.024	0.8214	45	0.73	1434	
	65	0.8927	63	0.97	0.829	0.016	0.8132	45	0.70	1676	
	66	0.8856	65	0.97	0.825	0.004	0.8210	45	0.69	1794	
	60	0.8843	63	0.96	0.873	-0.047	0.9200	45	0.75	1239	
	59	0.8816	60	0.96	0.856	0.017	0.8395	45	0.74	1317	
	67	0.8759	65	0.96	0.819	0.037	0.7824	45	0.67	2004	
	72	0.8701	67	0.96	0.809	0.010	0.7992	45	0.65	2172	
	71	0.8645	72	0.98	0.802	0.007	0.7957	45	0.63	2288	
	68	0.8599	67	0.96	0.807	-0.004	0.8111	45	0.64	2265	
	70	0.8538	68	0.96	0.781	0.026	0.7551	45	0.61	2472	
	69	0.8498	70	0.97	0.799	-0.018	0.8179	45	0.63	2346	
	74	0.8427	70	0.95	0.751	0.049	0.7019	45	0.56	2702	
	75	0.8371	74	0.97	0.761	-0.011	0.7721	45	0.56	2703	
	76	0.8314	75	0.97	0.751	0.010	0.7410	45	0.54	2771	
	78	0.8247	76	0.96	0.728	0.023	0.7055	45	0.51	2873	
	79	0.8173	78	0.96	0.707	0.022	0.6853	45	0.48	2954	
	73	0.8162	76	0.95	0.799	-0.092	0.8903	45	0.58	2613	
	80	0.8079	79	0.95	0.674	0.124	0.5504	45	0.43	3040	
	77	0.8044	76	0.95	0.747	-0.072	0.8190	45	0.50	2888	
	43	0.8042	66	0.93	0.801	-0.054	0.8554	80	0.68	1921	30
2	19										
	20	0.9788									
	18	0.9642	19	0.97	0.957	0.022	0.9350	20	0.95	100	
	17	0.9577	19	0.96	0.951	0.006	0.9454	18	0.95	104	
	16	0.9552	17	0.98	0.952	-0.000	0.9520	18	0.93	197	
	15	0.9463	16	0.96	0.928	0.023	0.9052	18	0.89	376	

	22	0.9382	20	0.95	0.918	0.010	0.9078	15	0.89	404	
	23	0.9336	22	0.96	0.920	-0.002	0.9214	15	0.87	511	
	24	0.9258	23	0.97	0.898	0.021	0.8770	15	0.83	701	
	28	0.9146	24	0.95	0.870	0.029	0.8409	15	0.79	931	
	29	0.9058	28	0.97	0.866	0.003	0.8629	15	0.78	987	
	27	0.8998	29	0.96	0.870	-0.004	0.8742	15	0.79	932	
	25	0.9006	27	0.96	0.905	-0.035	0.9394	15	0.82	748	
	26	0.9001	25	0.96	0.897	0.008	0.8891	15	0.83	710	
	32	0.8944	29	0.96	0.858	0.039	0.8184	15	0.75	1275	
	31	0.8901	32	0.97	0.860	-0.002	0.8615	15	0.75	1263	
	30	0.8839	31	0.95	0.838	0.022	0.8161	15	0.74	1374	
	33	0.8812	32	0.95	0.860	-0.022	0.8814	15	0.73	1402	
	13	0.8749	16	0.95	0.821	0.039	0.7819	30	0.69	1830	
	12	0.8687	13	0.97	0.814	0.007	0.8062	30	0.68	1901	
	64	0.8627	12	0.97	0.805	0.009	0.7964	30	0.67	1974	
	11	0.8573	12	0.96	0.804	0.001	0.8025	30	0.65	2193	
	14	0.8555	13	0.96	0.836	-0.032	0.8684	30	0.68	1889	
	10	0.8507	11	0.96	0.798	0.038	0.7597	30	0.63	2349	
	9	0.8452	11	0.95	0.783	0.015	0.7672	30	0.60	2501	
	21	0.8470	22	0.95	0.869	-0.087	0.9556	9	0.78	1022	
	41	0.8440	30	0.95	0.805	0.064	0.7418	9	0.63	2299	
	34	0.8384	33	0.95	0.767	0.039	0.7278	9	0.57	2635	
	35	0.8341	34	0.96	0.775	-0.008	0.7832	9	0.57	2665	
	36	0.8294	35	0.96	0.764	0.011	0.7535	9	0.55	2751	
	37	0.8242	36	0.96	0.748	0.016	0.7325	9	0.52	2837	
	39	0.8180	37	0.96	0.726	0.022	0.7035	9	0.49	2923	
	38	0.8129	37	0.96	0.733	-0.007	0.7403	9	0.50	2907	
	40	0.8055	39	0.94	0.687	0.046	0.6411	9	0.44	3030	
	8	0.7983	9	0.94	0.680	0.007	0.6738	40	0.39	3091	
	5	0.7901	8	0.95	0.650	0.030	0.6205	40	0.35	3125	
	4	0.7825	5	0.96	0.648	0.002	0.6461	40	0.33	3133	
	6	0.7774	4	0.96	0.687	-0.039	0.7263	40	0.37	3109	
	3	0.7704	5	0.96	0.641	0.047	0.5942	40	0.31	3144	
	1	0.7622	3	0.94	0.605	0.035	0.5702	40	0.26	3153	
	2	0.7556	1	0.97	0.626	-0.021	0.6467	40	0.28	3150	
	43	0.7566	21	0.94	0.777	-0.151	0.9275	40	0.60	2533	30
3	42										
	43	0.9393									
	21	0.9198	43	0.94	0.910	0.029	0.8808	42	0.88	416	4

POINTS NOT PLACED IN CLUSTERS
7 44

CONFIGURATION III


70
62
64
80
78
75
66
68
72
74
43
47
46
42
40
59,60
38
55
56
34
36
32
8
30
28
26
52
81
24
22
9
12
15
10
13
7
11
20
17,18
16
6
5
19
3
4
49
53
2
50
14
21
23
27
25
31
29
35
54
33
39
58
37
48
57
41
44
45
65
71
77
73
67
76
61
69
63
79
1

```
CLUSTER ANALYSIS OF RELATIVE SIMILARITIES

SAMPLE OF 80 POINTS IN TWO-SPACE

POINTS ARE CONSECUTIVELY NUMBERED FROM 1 TO  80

NUMBER OF POINTS IS  80   NUMBER OF SIMS IS   3160
NUMBER OF SIMS TO BE PUT IN CORE IS   3160
MEDIAN SIMILARITY IS 0.66272  AVERAGE SIMILARITY IS 0.63865

CLUSTERING TO BE AT 12 LEVELS OF RESOLUTION
BEGINNING AT 0.900 AND ENDING AT 0.442 BY STEPS OF 0.042

REASONS FOR CLUSTER TERMINATION

FLAG
COL.1=1  TOO LARGE A DROP IN AVERAGE LINKAGE
COL.2=2  TOO LARGE A DROP IN BEST LINK
COL.3=3  TOO LARGE A DECREASE IN THE DISTANCE OF NEW POINT
         FROM MOST DISTANT POINT IN THE CLUSTER
COL.4=4  POINT ALREADY INCLUDED IN A CLUSTER
```

FREQUENCIES OF SIMILARITIES AT 0.01 LEVELS

LEVEL	NUMBER	TOTAL,	REMAINDER	HISTOGRAM
1.00	12	12	3148	***
0.99	11	23	3137	**
0.98	27	50	3110	******
0.97	27	77	3083	******
0.96	20	97	3063	*****
0.95	38	135	3025	*********
0.94	28	163	2997	*******
0.93	24	187	2973	******
0.92	41	228	2932	**********
0.91	34	262	2898	********
0.90	24	286	2874	******
0.89	32	318	2842	********
0.88	17	335	2825	****
0.87	37	372	2788	*********
0.86	28	400	2760	*******
0.85	22	422	2738	*****
0.84	31	453	2707	*******
0.83	17	470	2690	****
0.82	28	498	2662	*******
0.81	31	529	2631	*******
0.80	29	558	2602	*******
0.79	35	593	2567	********
0.78	37	630	2530	*********
0.77	55	685	2475	*************
0.76	68	753	2407	*****************
0.75	49	802	2358	************
0.74	73	875	2285	******************
0.73	80	955	2205	********************
0.72	90	1045	2115	**********************
0.71	125	1170	1990	*******************************
0.70	140	1310	1850	***********************************
0.69	94	1404	1756	***********************
0.68	102	1506	1654	*************************
0.67	88	1594	1566	**********************
0.66	82	1676	1484	********************
0.65	77	1753	1407	*******************
0.64	66	1819	1341	****************
0.63	49	1868	1292	************
0.62	58	1926	1234	**************
0.61	35	1961	1199	********
0.60	44	2005	1155	***********
0.59	27	2032	1128	******
0.58	31	2063	1097	*******
0.57	41	2104	1056	**********
0.56	26	2130	1030	******
0.55	40	2170	990	**********
0.54	45	2215	945	***********
0.53	28	2243	917	*******
0.52	35	2278	882	********
0.51	37	2315	845	*********

0.50	31	2346	814	*******
0.49	58	2404	756	**************
0.48	39	2443	717	*********
0.47	58	2501	659	**************
0.46	41	2542	618	**********
0.45	71	2613	547	*****************
0.44	55	2668	492	*************
0.43	63	2731	429	***************
0.42	94	2825	335	***********************
0.41	101	2926	234	*************************
0.40	72	2998	162	******************
0.39	57	3055	105	**************
0.38	35	3090	70	********
0.37	22	3112	48	*****
0.36	5	3117	43	*
0.35	8	3125	35	**
0.34	8	3133	27	**
0.33	0	3133	27	
0.32	2	3135	25	*
0.31	0	3135	25	
0.30	1	3136	24	*
0.29	0	3136	24	
0.28	2	3138	22	*
0.27	1	3139	21	*
0.26	3	3142	18	*
0.25	0	3142	18	
0.24	1	3143	17	*
0.23	1	3144	16	*
0.22	0	3144	16	
0.21	3	3147	13	*
0.20	5	3152	8	*
0.19	4	3156	4	*
0.18	1	3157	3	*
0.17	2	3159	1	*
0.16	0	3159	1	
0.15	1	3160	-0	*
0.14	0	3160	-0	
0.13	0	3160	-0	
0.12	0	3160	-0	
0.11	0	3160	-0	
0.10	0	3160	-0	
0.09	0	3160	-0	
0.08	0	3160	-0	
0.07	0	3160	-0	
0.06	0	3160	-0	
0.05	0	3160	-0	
0.04	0	3160	-0	
0.03	0	3160	-0	
0.02	0	3160	-0	
0.01	0	3160	-0	

CLUSTERS AT RESOLUTION LEVEL 2 RESOL1 IS 0.858
RESOL2 IS 0.879

CLUSTR	AND MEMBER	AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	17										
	18	1.0000									
	24	0.9484	17	0.92	0.923	0.077	0.8452	18	0.92	183	1200
2	59										
	60	1.0000									
	43	0.9937	60	0.99	0.991	0.009	0.9812	59	0.99	9	
	47	0.9813	59	0.97	0.969	0.022	0.9472	43	0.97	58	
	40	0.9694	43	0.96	0.951	0.017	0.9340	47	0.93	168	
	56	0.9594	40	0.98	0.939	0.012	0.9275	47	0.91	236	
	36	0.9498	56	0.98	0.926	0.014	0.9121	47	0.88	312	
	66	0.9335	47	0.94	0.884	0.041	0.8429	36	0.82	467	1000
3	73										
	76	0.9934									
	79	0.9804	73	0.98	0.974	0.019	0.9545	76	0.97	40	
	63	0.9658	79	0.97	0.951	0.023	0.9282	76	0.94	133	
	71	0.9599	63	0.97	0.951	-0.000	0.9512	76	0.94	137	
	69	0.9446	63	0.95	0.914	0.037	0.8769	76	0.89	302	
	67	0.9383	69	0.97	0.923	-0.008	0.9310	76	0.89	294	
	77	0.9278	67	0.96	0.896	0.026	0.8704	76	0.85	406	
	65	0.9207	77	0.99	0.896	0.001	0.8947	76	0.83	438	
	61	0.9125	65	0.96	0.880	0.016	0.8638	76	0.81	516	
	45	0.9024	76	0.94	0.857	0.023	0.8346	61	0.75	722	1000
4	74										
	75	0.9934									
	80	0.9804	74	0.98	0.974	0.019	0.9545	75	0.97	39	
	64	0.9658	80	0.97	0.951	0.023	0.9282	75	0.94	132	
	72	0.9599	64	0.97	0.951	-0.000	0.9512	75	0.94	138	
	70	0.9446	64	0.95	0.914	0.037	0.8769	75	0.89	301	
	68	0.9383	70	0.97	0.923	-0.008	0.9310	75	0.89	295	
	78	0.9278	68	0.96	0.896	0.026	0.8704	75	0.85	407	
	66	0.9207	78	0.99	0.896	0.001	0.8947	75	0.83	439	
	62	0.9125	66	0.96	0.880	0.016	0.8638	75	0.81	515	
	46	0.9024	75	0.94	0.857	0.023	0.8346	62	0.75	721	1000
5	9										
	12	0.9934									
	15	0.9804	12	0.98	0.974	0.019	0.9545	9	0.97	45	

	10	0.9683	15	0.97	0.956	0.018	0.9384	9	0.95	106	
	11	0.9656	10	0.99	0.961	-0.005	0.9668	9	0.94	125	
	16	0.9637	11	0.99	0.960	0.002	0.9583	9	0.94	136	
	6	0.9611	16	0.99	0.955	0.005	0.9497	12	0.93	160	
	5	0.9558	6	0.97	0.940	0.015	0.9249	12	0.91	239	
	4	0.9497	5	0.98	0.928	0.011	0.9170	9	0.89	293	
	13	0.9475	10	0.97	0.939	-0.010	0.9488	4	0.91	200	
	14	0.9399	4	0.96	0.906	0.033	0.8725	9	0.86	383	
	7	0.9375	6	0.96	0.925	-0.020	0.9453	14	0.88	317	
	2	0.9328	4	0.95	0.907	0.018	0.8893	12	0.87	358	
	3	0.9275	2	0.96	0.895	0.012	0.8832	13	0.85	403	
	8	0.9212	9	0.94	0.880	0.015	0.8652	14	0.81	511	
	47	0.8949	8	0.82	0.711	0.169	0.5416	14	0.63	1839	1204
6	44										
	58	0.9934									
	48	0.9774	58	0.97	0.969	0.024	0.9454	44	0.97	61	
	39	0.9629	44	0.96	0.948	0.021	0.9274	48	0.93	169	
	35	0.9439	39	0.95	0.915	0.033	0.8824	48	0.88	311	
	65	0.9253	48	0.94	0.888	0.027	0.8610	35	0.82	468	4
7	41										
	57	0.9934									
	37	0.9738	57	0.97	0.964	0.029	0.9347	41	0.96	70	
	45	0.9626	41	0.97	0.951	0.013	0.9388	37	0.93	167	
	33	0.9448	37	0.95	0.918	0.033	0.8846	45	0.88	310	
	54	0.9365	33	0.99	0.920	-0.002	0.9216	45	0.87	342	
	76	0.9198	45	0.94	0.878	0.041	0.8369	54	0.81	492	1004
8	21										
	50	0.9906									
	25	0.9776	21	0.97	0.971	0.020	0.9515	50	0.97	50	
	29	0.9659	25	0.97	0.954	0.017	0.9372	50	0.94	126	
	49	0.9586	50	0.97	0.948	0.006	0.9413	29	0.91	210	
	19	0.9415	49	0.95	0.907	0.040	0.8670	29	0.86	360	
	20	0.9328	19	0.99	0.911	-0.004	0.9148	29	0.85	395	
	22	0.9169	20	0.93	0.869	0.042	0.8274	29	0.81	518	1000
9	28										
	52	0.9906									
	24	0.9812	52	0.98	0.977	0.014	0.9624	28	0.97	41	
	32	0.9712	28	0.97	0.961	0.015	0.9457	24	0.95	115	
	17	0.9418	24	0.92	0.898	0.063	0.8343	32	0.87	336	1204
10	22										
	51	0.9852									
	26	0.9807	51	0.99	0.978	0.007	0.9718	22	0.97	43	

	30	0.9704	26	0.97	0.960	0.018	0.9418	22	0.95	116	
	20	0.9439	22	0.93	0.904	0.056	0.8483	30	0.87	328	1204
11	23										
	53	0.9852									
	27	0.9765	53	0.97	0.972	0.013	0.9593	23	0.97	42	
	31	0.9657	27	0.97	0.955	0.017	0.9373	53	0.95	117	
	35	0.9314	31	0.91	0.880	0.075	0.8052	53	0.86	374	1204
12	34										
	55	0.9812									
	38	0.9612	34	0.95	0.951	0.030	0.9213	55	0.95	99	
	42	0.9452	38	0.96	0.929	0.022	0.9069	55	0.91	223	
	46	0.9327	42	0.97	0.914	0.015	0.8992	55	0.88	325	
	75	0.9135	46	0.94	0.875	0.039	0.8358	55	0.82	464	1004

POINTS NOT PLACED IN CLUSTERS

1

CLUSTERS AT RESOLUTION LEVEL 11 RESOL1 IS 0.483

RESOL2 IS 0.692

CLUSTR AND	MEMBER	AVGSIM OF CLUSTR	OTU-BEST LINK	SIM-BEST LINK	AVGOF NEW LINKS	DROP IN AVG	DIFF	FAR OTU	FAR SIM	RANK FAR SIM	FLAG
1	17										
	18	1.0000									
	24	0.9484	17	0.92	0.923	0.077	0.8452	18	0.92	183	
	52	0.9391	24	0.98	0.930	-0.007	0.9372	17	0.90	257	
	28	0.9387	52	0.99	0.938	-0.008	0.9463	17	0.90	270	
	32	0.9339	28	0.97	0.924	0.014	0.9105	18	0.87	337	
	36	0.9102	32	0.91	0.851	0.073	0.7774	18	0.78	574	
	56	0.8954	36	0.98	0.851	-0.000	0.8510	17	0.76	669	
	40	0.8843	56	0.98	0.846	0.005	0.8406	17	0.74	815	
	43	0.8729	40	0.96	0.827	0.019	0.8086	17	0.70	1133	
	60	0.8676	43	0.99	0.844	-0.017	0.8604	18	0.71	1118	
	59	0.8660	60	1.00	0.858	-0.014	0.8722	18	0.71	1117	
	47	0.8623	59	0.97	0.842	0.016	0.8260	18	0.68	1424	
	66	0.8524	47	0.94	0.793	0.049	0.7436	17	0.62	1880	
	78	0.8445	66	0.99	0.793	-0.000	0.7935	18	0.61	1944	
	62	0.8408	66	0.96	0.815	-0.022	0.8374	17	0.61	1908	
	68	0.8340	78	0.96	0.782	0.033	0.7494	17	0.58	2048	
	70	0.8271	68	0.97	0.773	0.010	0.7627	18	0.55	2132	
	64	0.8186	70	0.95	0.746	0.027	0.7191	17	0.52	2245	
	72	0.8139	64	0.97	0.772	-0.026	0.7983	17	0.55	2150	
	80	0.8071	64	0.97	0.742	0.030	0.7118	17	0.50	2319	
	74	0.8015	80	0.98	0.746	-0.004	0.7496	18	0.50	2317	
	75	0.7971	74	0.99	0.751	-0.005	0.7565	17	0.49	2339	
	46	0.7906	75	0.94	0.719	0.032	0.6874	18	0.47	2440	
	42	0.7833	46	0.97	0.699	0.020	0.6793	17	0.44	2581	
	38	0.7752	42	0.96	0.678	0.021	0.6570	18	0.42	2706	
	34	0.7662	38	0.95	0.654	0.024	0.6298	17	0.40	2889	
	55	0.7584	34	0.98	0.656	-0.002	0.6586	18	0.39	3035	
	30	0.7491	34	0.91	0.623	0.033	0.5901	18	0.39	2991	
	26	0.7404	30	0.97	0.619	0.004	0.6150	17	0.39	3005	
	51	0.7325	26	0.99	0.618	0.001	0.6171	17	0.38	3067	
	22	0.7254	51	0.99	0.619	-0.001	0.6194	18	0.37	3082	
	20	0.7159	22	0.93	0.568	0.050	0.5180	18	0.34	3131	
	19	0.7075	20	0.99	0.574	-0.006	0.5794	18	0.34	3132	
	49	0.7004	19	0.95	0.582	-0.008	0.5909	17	0.37	3101	
	50	0.6939	49	0.97	0.583	-0.001	0.5840	18	0.38	3065	
	21	0.6881	50	0.99	0.587	-0.003	0.5901	17	0.37	3087	
	25	0.6830	21	0.97	0.592	-0.006	0.5981	17	0.39	3027	
	29	0.6781	25	0.97	0.586	0.007	0.5792	18	0.39	3012	
	54	0.6714	29	0.92	0.545	0.041	0.5045	62	0.38	3058	
	33	0.6657	54	0.99	0.555	-0.010	0.5644	62	0.38	3050	
	37	0.6600	33	0.95	0.545	0.010	0.5354	62	0.37	3089	
	57	0.6545	37	0.97	0.543	0.002	0.5404	62	0.36	3106	
	41	0.6500	57	0.99	0.554	-0.011	0.5650	62	0.37	3102	

	45	0.6459	41	0.97	0.559	-0.005	0.5647	62	0.37	3074	
	76	0.6413	45	0.94	0.540	0.020	0.5202	62	0.36	3114	
	73	0.6374	76	0.99	0.549	-0.009	0.5583	62	0.36	3110	
	79	0.6333	73	0.98	0.541	0.008	0.5324	62	0.35	3122	
	63	0.6295	79	0.97	0.539	0.001	0.5377	70	0.35	3120	
	71	0.6272	63	0.97	0.571	-0.032	0.6030	70	0.38	3054	
	69	0.6235	63	0.95	0.533	0.038	0.4958	70	0.34	3128	
	67	0.6212	69	0.97	0.564	-0.030	0.5944	70	0.37	3091	
	77	0.6188	67	0.96	0.558	0.006	0.5527	70	0.37	3100	
	65	0.6169	77	0.99	0.566	-0.007	0.5727	70	0.37	3078	
	61	0.6142	65	0.96	0.545	0.020	0.5245	70	0.34	3125	
	48	0.6129	65	0.94	0.576	-0.031	0.6062	80	0.40	2968	
	58	0.6115	48	0.97	0.573	0.002	0.5712	55	0.38	3042	
	44	0.6103	58	0.99	0.577	-0.004	0.5814	55	0.38	3059	
	39	0.6091	44	0.96	0.575	0.002	0.5736	55	0.37	3085	
	35	0.6080	39	0.95	0.576	-0.001	0.5773	55	0.37	3093	
	31	0.6073	35	0.91	0.587	-0.011	0.5977	20	0.39	3023	
	27	0.6070	31	0.97	0.595	-0.008	0.6038	20	0.39	3036	
	53	0.6068	27	0.97	0.602	-0.007	0.6092	20	0.38	3041	
	23	0.6066	53	0.99	0.599	0.003	0.5959	20	0.37	3090	
	8	0.6091	47	0.82	0.687	-0.088	0.7752	69	0.56	2103	230
2	9										
	12	0.9934									
	15	0.9804	12	0.98	0.974	0.019	0.9545	9	0.97	45	
	10	0.9683	15	0.97	0.956	0.018	0.9384	9	0.95	106	
	11	0.9656	10	0.99	0.961	-0.005	0.9668	9	0.94	125	
	16	0.9637	11	0.99	0.960	0.002	0.9583	9	0.94	136	
	6	0.9611	16	0.99	0.955	0.005	0.9497	12	0.93	160	
	5	0.9558	6	0.97	0.940	0.015	0.9249	12	0.91	239	
	4	0.9497	5	0.98	0.928	0.011	0.9170	9	0.89	293	
	13	0.9475	10	0.97	0.939	-0.010	0.9488	4	0.91	200	
	14	0.9399	4	0.96	0.906	0.033	0.8725	9	0.86	383	
	7	0.9375	6	0.96	0.925	-0.020	0.9453	14	0.88	317	
	2	0.9328	4	0.95	0.907	0.018	0.8893	12	0.87	358	
	3	0.9275	2	0.96	0.895	0.012	0.8832	13	0.85	403	
	8	0.9212	9	0.94	0.880	0.015	0.8652	14	0.81	511	
	47	0.8949	8	0.82	0.711	0.169	0.5416	14	0.63	1839	204

POINTS NOT PLACED IN CLUSTERS
1